



Morphologie et architecture des interfaces de communication de l'information scientifique et technique dans un environnement multilingue: le contexte arabo-latin

Mokhtar Ben Henda

► To cite this version:

Mokhtar Ben Henda. Morphologie et architecture des interfaces de communication de l'information scientifique et technique dans un environnement multilingue: le contexte arabo-latin. Interface homme-machine [cs.HC]. Université Michel de Montaigne - Bordeaux III, 1999. Français. NNT: . tel-00006373

HAL Id: tel-00006373

<https://theses.hal.science/tel-00006373>

Submitted on 13 Sep 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

UNIVERSITÉ MICHEL DE MONTAIGNE-BORDEAUX III

UFR Sciences de l'Information, de la Communication et des Arts

**MORPHOLOGIE ET ARCHITECTURE
DES INTERFACES DE COMMUNICATION
DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE
DANS UN ENVIRONNEMENT MULTILINGUE**

Le contexte arabo-latin

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ

Discipline : Sciences de l'Information et de la Communication

Présentée et soutenue publiquement par :

Mokhtar BEN HENDA

Sous la direction du Professeur Roland DUCASSE

Jury :

Richard BOUCHÉ, *Professeur*
Khalifa CHATER, *Professeur*
Roland DUCASSE, *Professeur*
Albert MAILLES, *Maître de Conférences*

Février 1999

DEDICACE

À vous

MON PÈRE
nonagénaire,

En mémoire d'une question existentielle majeure que vous m'avez posée un mois de mai 1970 et que je n'oserais jamais vous rappeler!

... parce qu'entre autres je n'y ai pas encore de réponse !

... une question qui a semé en moi le premier germe de la réflexion.

À vous qui m'avez aiguillé sur les voies du savoir, qui m'avez toujours donné confiance et initié au langage de l'esprit.

À vous

MA MÈRE
octogénaire

En mémoire de la gargoulette et du morceau de charbon d'il y a 35 ans environ ...

Nous sommes peut-être les seuls à nous en souvenir.

À vous

Mes frères et sœurs

Vous avez toujours voulu gâter votre petit frère. J'espère toujours ne pas vous décevoir.

À toi

Ma Dulcinée virtuelle
Puisses tu un jour devenir réelle !

Je vous dédie à tous cet humble travail comme signe de gratitude et de reconnaissance éternelle.

REMERCIEMENTS

A celui qui a été patient face à mes cavales et mes périodes « d'ermitage » prolongées, à celui qui m'a toujours fait confiance et aidé à retrouver le chemin de la réussite scientifique, à mon encadreur Monsieur le Professeur **Roland DUCASSE**, veuillez accepter mes vifs remerciements pour toutes les facilités que vous avez mises à ma disposition tout au long de ces années durant lesquelles nous avons travaillé ensemble.

À tous les amis et collègues du Centre d'Étude des Medias et du Groupe de recherche expérimentale sur les systèmes informatisés de la communication, veuillez trouver ici l'expression de mes meilleurs remerciements pour m'avoir accueilli si chaleureusement parmi vous, pour m'avoir soutenu et encouragé tout au long de ce travail.

À deux amies et sœurs qui me sont très chères : **Chehrazède** et **Salwa**, soyez toujours la lumière qui guidera mon esprit sur les voies de la vertu, de l'amitié et de la fraternité pures et sincères.

À tous mes étudiants de l'Institut Supérieur de Documentation de Tunis pour leur sympathie et soutien continus.

Aux membres du Jury qui ont daigné évaluer ce travail et subir les tourments linguistiques et stylistiques que je leur inflige, veuillez trouver ici ma reconnaissance et mes remerciements les plus sincères.

Table des matières

CHAPITRE PREMIER

ÉTAT DE LA QUESTION

1. INTRODUCTION	1
1.1 UNE SOCIÉTÉ EN MUTATION : CONVERGENCE OU DIVERGENCE DE MODÈLES ?	1
1.2 LA RÉVOLUTION DE L'IMMATÉRIEL : UNE NOUVELLE DONNÉE SOCIO-ÉCONOMIQUE	4
2. PROBLÉMATIQUE	19
2.1 PREMIER AXE DE RECHERCHE	20
2.2 DEUXIÈME AXE DE RECHERCHE	21
3. OBJECTIFS DE L'ÉTUDE	22
4. MÉTHODOLOGIE DE TRAVAIL	23
5. PRÉCISIONS TERMINOLOGIQUES	24
6. REVUE DE LA LITTÉRATURE	28
6.1 ÉTUDES ET PUBLICATIONS	28
6.2 DES SITES WEB	39

CHAPITRE DEUXIÈME

LES ASPECTS DE LA COMMUNICABILITÉ DE L'INFORMATION AU DIAPASON DES DISCIPLINES CONNEXES

1. LES THÉORIES LINGUISTIQUES ET LA SEMIOLOGIE DU LANGAGE : DE L'ABSTRACTION PLATONICIENNE AU STRUCTURALISME SAUSSURIEN	44
1.1 À LA RECHERCHE D'UNE DÉFINITION DE LA LANGUE ET DU LANGAGE	45
1.2 LES THÉORIES DU LANGAGE : ENTRE LA PERCEPTIBILITÉ DU SIGNE ET L'INTELLIGIBILITÉ DU SENS	46
1.3 LE PARADIGME DU « SIGNIFIANT/SIGNIFIÉ »	48
1.4 SEMIOTIQUE ET SYMBOLISME DANS LA THÉORIE LINGUISTIQUE	50
1.5 LA NOTION DE « SYSTÈME » DANS LA PHILOSOPHIE LINGUISTIQUE	51
1.6 LES PARADIGMES DE LA COMMUNICATION LANGAGIÈRE DANS LES THÉORIES DE L'INFORMATION ET DE LA COMMUNICATION	52
1.7 DE L'ORALITÉ À L'ÉCRIT : ENTRE LA MÉTAPHYSIQUE ET LA MATÉRIALITÉ DU SENS	58
1.8 ÉCRITURE ÉLECTRONIQUE : ENTRE LA DOUBLURE DE CODAGE ET LA VARIABILITÉ DE TRAITEMENT	59
2. LA THÉORIE DU SIGNAL	60
2.1 L'ÉNERGIE ÉLECTRIQUE COMME SOURCE DE SIGNAL	61
2.2 LE TRAITEMENT DU SIGNAL	62
3. LA REPRÉSENTATION DES DONNÉES ET LE CODAGE DE L'INFORMATION	63
3.1 LE CODAGE BINAIRE DANS LES SYSTÈMES DE TRAITEMENT ET D'ÉCHANGE DES DONNÉES	65
3.2 LE CODAGE BINAIRE : LES AVATARS DU NUMÉRIQUE	68
4. L'ALGORITHMIQUE	71
4.1. ÉTYMOLOGIE DU CONCEPT	71
4.2 DÉFINITIONS DU CONCEPT	72
4.3 STRUCTURE ALGORITHMIQUE	72
4.4 NATURE ET DOMAINES D'APPLICATION	75
5. CONCLUSION	78

CHAPITRE TROISIÈME

CODAGE, NORMALISATION ET TRANSPARENCE LINGUISTIQUE

1. L'ASCII : L'HISTOIRE D'UN STANDARD DEVENU UN MODELE INTERNATIONAL	80
1.1 LA « BIOGRAPHIE D'UN DINOSAURE »	81
1.2 LA STRUCTURE BINAIRE DE L'ASCII	82
1.3 L'IRV, L'EMBRYON MULTILINGUE DE L'ASCII	85
2. L'EXTENSION DES CODES : UN TREMPLIN VERS UN CONTEXTE MULTILINGUE DE CODAGE ET D'ECHANGE DE L'INFORMATION	87
2.1 STRUCTURE D'UNE TABLE DES CODES ETENDUS (A 8 BITS)	88
2.2 ISO 2022 : UNE NORME D'EXTENSION DE CODES	90
2.3 CONCLUSION	94
3. CODAGE, NORMALISATION ET MULTILINGUISME ARABE/LATIN	95
3.1 L'ASMO 449 : UN DEBUT DE NORMALISATION	95
3.2 ASMO 662 = ASCII ETENDU (LATIN 2)	99
3.3 ASMO 708 (ISO/IEC 8859-6 = ECMA 114)	101
3.4 LES JEUX DE CARACTERES CODES SUR PLUSIEURS OCTETS	105
4. LE CODAGE DES CARACTERES DANS LES SYSTEMES OUVERTS ET DISTRIBUES : LE CAS INTERNET	127
4.1 LE SCHEMA DE FONCTIONNEMENT DU WORLD WIDE WEB	128
4.2 LA DIMENSION MULTILINGUE DANS LES PROTOCOLES INTERNET	131
5. CONCLUSION	224

CHAPITRE QUATRIEME

CATALYSEURS ERGONOMIQUES DES INTERFACES HOMME-MACHINE DANS LES SYSTEMES D'INFORMATION MULTILINGUES ARABE/LATIN

1. LA COMPOSITION DU TEXTE ÉCRIT : ENTRE LE SENS DU CARACTERE ET LA VISIBILITÉ DU GLYPHE	225
2. LES COMPLEXITES D'USAGE DE L'ÉCRITURE ARABE	228
2.1 LES CODES DES CARACTERES ET LES LACUNES DES NORMES DE CODIFICATION	228
2.2 LE TRI ALPHABETIQUE : LA DIVERSITE DES CONTEXTES ET LA VARIATION DES ALGORITHMES	241
2.3 LA BIDIRECTIONNALITE	269
3. LES CHIFFRES, LA BIDIRECTIONNALITÉ ET L'INTERNATIONALISATION : LA CONFLICTIVE CULTURELLE	283
3.1 LES SYSTEMES NUMERIQUES ET LA REPRESENTATION GRAPHIQUE	284
3.2 L'UNIDIRECTIONNALITE ALGEBRIQUE ET LA BIDIRECTIONALITE ALPHANUMERIQUE	296

CONCLUSION GÉNÉRALE

Erreur ! Signet non défini.

BIBLIOGRAPHIE UTILISEE

Erreur ! Signet non défini.

BIBLIOGRAPHIE RECOMMANDEE

Erreur ! Signet non défini.

ANNEXE A	337
ANNEXE B	346
ANNEXE C	348
ANNEXE D1	354
ANNEXE D2	354
ANNEXE E	365
ANNEXE F	367
ANNEXE G	370
ANNEXE H	373
ANNEXE I	377

Liste des tableaux

TABLEAU 1 : LES CRITERES DE DEVELOPPEMENT D'UNE SOCIETE D'INFORMATION	6
TABLEAU 2 : STRATES DE LA STRUCTURE LINGUISTIQUE D'UNE ENTITE MULTILINGUE	21
TABLEAU 3: TABLEAU DE VALIDATION DES AXIOMES MATHEMATIQUES DE BOOLE PAR DES ELEMENTS BINAIRES.....	66
TABLEAU 4 : LA REPRESENTATION BINAIRE PAR SYSTEME PONDERE	70
TABLEAU 5 : LA REPRESENTATION BINAIRE PAR SYSTEME REFLECHI	70
TABLEAU 6: PRINCIPALES NORMES DE CODAGE.....	80
TABLEAU 7 : POSITIONS ET POIDS DES COMBINAISONS BINAIRES A 7 BITS.....	83
TABLEAU 8 : LA TABLE ASCII = ISO 646.....	83
TABLEAU 9 : LES REFERENCES NATIONALES ISO/IEC 646	86
TABLEAU 10 : LES CARACTERES OPTIONNELS DE LA VERSION DE REFERENCE IRV DE LA NORME ISO 646	86
TABLEAU 11 : LES 4 ZONES DE CARACTERES D'UNE TABLE A 8 BITS	88
TABLEAU 12 : LA STRUCTURE ETENDUE D'UNE TABLE A 8 BITS (ASCII ETENDU)	89
TABLEAU 13 : LES 32, 94 ET 96 POSITIONS UTILISEES PAR LE JEU C0 ET LES 4 JEUX G0/G3.....	91
TABLEAU 14 : CODAR-1	96
TABLEAU 15 : CODAR-2	96
TABLEAU 16 : CODAR-U.....	97
TABLEAU 17 : LA VERSION DEFINITIVE CODAR-U/FD	98
TABLEAU 18 : LE JEU DE CARACTERES ARABE (ASMO 449)	99
TABLEAU 19 : LA NORME ASMO 662.....	100
TABLEAU 20 : LA NORME ASMO 708 (ISO/IEC 8856-6 / ECMA 114)	101
TABLEAU 21 : LA SERIE DE NORMES ISO/IEC 8859-x	103
TABLEAU 22 : PAIRES DES CARACTERES SYMETRIQUES ARABES/LATINS	105
TABLEAU 23 : LE BLOC UNICODE 0000-00FF	109
TABLEAU 24 : NORMES ET STANDARDS COMPATIBLES AVEC UNICODE	113
TABLEAU 25 : LA VALEUR BINAIRE DU CARACTERE A SELON PLUSIEURS STANDARDS	114
TABLEAU 26 : TRANSFORMATION UCS-4 / UTF-8	116
TABLEAU 27 : REPARTITION DES CARACTERES UNICODE SELON LA DIRECTION DE L'ECRITURE	126
TABLEAU 28 : NUANCES ENTRE MARQUES « LRM » ET « RTL » DANS LE TRAITEMENT LOGIQUE DES URL	154
TABLEAU 29 : LES VALEURS DU PARAMETRE « CHARSET ».....	157
TABLEAU 30 : DEGRES DE TRAITEMENT DE LA LANGUE ARABE PAR LES NAVIGATEURS INTERNET MULTILINGUES...	204
TABLEAU 31 : DES GLYPHES / CARACTERES DANS LES TABLES DES CODES ARABES	229
TABLEAU 32 : LES 28 LETTRES DE L'ALPHABET ARABE	230
TABLEAU 33 : PAGE DE CODES 1256 SOUS ARABIC MS WINDOWS 3.11	232
TABLEAU 34 : LES CARACTERES ARABES ET LEURS FORMES CONTEXTUELLES CONFORMES A ASMO 969.....	233
TABLEAU 35 : LE JEU MINIMAL DES FORMES DE CARACTERES ARABES SELON ASMO 969	234
TABLEAU 36 : TAUX DE FREQUENCE DES CARACTERES EN ARABE STANDARD	237
TABLEAU 37 : LA REPARTITION DES BLOCS DE CARACTERES GRAPHIQUE DANS LA TABLE ISO 646	248
TABLEAU 38 : LA REPARTITION DES BLOCS DE CARACTERES GRAPHIQUES DANS UNE TABLE LATIN/ARABE A 8 BITS	249
TABLEAU 39 : COMPARAISON DE L'ORDRE DE TRI ENTRE LA FONCTION « SORT » DE MS-DOS ET MS-WORD 97...	251
TABLEAU 40 : LA « KASHIDA » LE « LAM-ALEF » DANS UN TRI EN LANGUE ARABE SOUS ARABIC WORD 6 ET 97 ..	256
TABLEAU 41 : RENDU DE TRI MULTILINGUE SOUS ARABIC MS-DOS EN MML	258
TABLEAU 42 : TRI MIXTE ARABE-LATIN DES CARACTERES SPECIAUX SOUS ARABIC MS-WORD 97.....	261
TABLEAU 43 : TRI MIXTE ARABE-LATIN DES CARACTERES ALPHABETIQUES SOUS ARABIC MS-WORD 97	262
TABLEAU 44 : L'EFFET DE SWAPPING SOUS ARABIC MS DOS	278
TABLEAU 45 : EFFET DE SWAPPING SOUS MS EXCEL	279
TABLEAU 46 : EFFET DE SWAPPING SUR LES CHIFFRES EN MMA	283
TABLEAU 47 : LE SYSTEME DECIMAL A BASE 10.....	288
TABLEAU 48 : VALEURS POSITIONNELLES DANS UN SYSTEME DECIMAL	289
TABLEAU 49 : MECANISME DE CALCUL DU SYSTEME NUMERIQUE DECIMAL	289
TABLEAU 50 : VALEURS POSITIONNELLES DANS UN SYSTEME BINAIRE	289
TABLEAU 51 : MECANISME DE CALCUL DU SYSTEME NUMERIQUE BINAIRE	290
TABLEAU 52 : MECANISME INCREMENTIEL DES CHIFFRES	297
TABLEAU 53 : STRUCTURES MORPHOLOGIQUES DES CHIFFRES EN ALLEMAND, ARABE, FRANÇAIS ET ANGLAIS	305

Liste des figures

FIGURE 1 : LE SCHEMA DE SAUSSURE SUR LA COMMUNICATION LANGAGIERE	49
FIGURE 2 : SCHEMA DE SHANNON ET WEAVER, 1947	53
FIGURE 3 : SCHEMA DE RETROACTION DE BERTALANFFY	54
FIGURE 4 : LE SCHEMA DE BÜHLER.....	57
FIGURE 5 : LE SCHEMA DE COMMUNICATION ET LES FONCTIONS DU LANGAGE DANS LE MODELE JAKOBSON	57
FIGURE 6: SYNTHÈSE DES MODELES REIX RELATIFS AUX PRINCIPES DE LA REPRÉSENTATION DE L'INFORMATION.....	60
FIGURE 7 : L'ÉVOLUTION DES COUCHES INTERFACES INFORMATION/HOMME	65
FIGURE 8 : MODELE D'ANALYSE DE L'ENTROPIE DE SHANNON.....	66
FIGURE 9 : EXEMPLE DE CODAGE EN BINAIRE ALPHABÉTIQUE SELON LE MODELE BACON.....	67
FIGURE 10 : L'ARCHITECTURE INFORMATIQUE DE VON NEUMANN	69
FIGURE 11 : STRUCTURE ITERATIVE A CONDITION PRE-OPERATOIRE	73
FIGURE 12 : STRUCTURE ITERATIVE A CONDITION POST-OPERATOIRE.....	73
FIGURE 13 : STRUCTURE ALTERNATIVE	74
FIGURE 14 : STRUCTURE ALTERNATIVE COMBINÉE	74
FIGURE 15 : STRUCTURE GÉNÉRALE D'UNE TABLE DE DÉCISION.....	75
FIGURE 16 : SOUBASSEMENTS DE L'INTERACTIVITÉ HOMME-MACHINE.....	78
FIGURE 17 : LES ÉLÉMENTS DE CODE D'UN JEU DE CARACTÈRES	91
FIGURE 18 : EXTENSION DE CODE DANS UN ENVIRONNEMENT 7 BITS.....	93
FIGURE 19 : EXTENSION DE CODE DANS UN ENVIRONNEMENT 8 BITS.....	94
FIGURE 20 : STRUCTURE DES ESPACES DE CODAGE ISO/IEC 10646	109
FIGURE 21 : STRUCTURE D'UN GROUPE ISO/IEC 10646.....	110
FIGURE 22 : LES ZONES DU BMP ISO/IEC 10646.....	111
FIGURE 23 : DEDOUBLEMENT DE TAILLE POUR UN MEME FICHIER DU FORMAT ASCII AU FORMAT UNICODE.....	113
FIGURE 24: SAUVEGARDE EN FORMAT UNICODE SOUS MS WORD 97.....	119
FIGURE 25 : MODE DE CONVERSION CODE/GLYPHE PAR PAGES DE CODES ET DE FONTE	120
FIGURE 26 : MODE DE CONVERSION CODE/GLYPHE PAR FONTES UNICODE	120
FIGURE 27 : UNIFICATION DES FONTES UNICODE	121
FIGURE 28 : SCHEMA DE FONCTIONNEMENT DU WWW.....	129
FIGURE 29 : CHOIX DES PROTOCOLES SOUS L'ÉDITEUR HTMLÉD32 DE 2 ^E GÉNÉRATION.....	130
FIGURE 30 : ARTICULATION ENTRE PROTOCOLES INTERNET.....	131
FIGURE 31 : STRUCTURE D'UNE ADRESSE URL	133
FIGURE 32 : NIVEAUX DE DÉFINITION DES CARACTÈRES URL	134
FIGURE 33 : RAPPORTS URLs, URNs & URIs.....	136
FIGURE 34 : FONCTIONNEMENT TYPIQUE D'UN SERVICE DE NOMS DE DOMAINES	137
FIGURE 35 : INSTALLATION ET FONCTIONNEMENT STANDARDS DES DNS	138
FIGURE 36 : CORRESPONDANCE ENTRE URLs ET URNs VIA DNS	140
FIGURE 37 : URL STATIQUE EN ASCII POUR UN DOCUMENT EN CP ARABE WINDOWS 1256.....	141
FIGURE 38 : ACCÈS UNILINGUE VIA CODES DE CARACTÈRES UNIFIÉS.....	143
FIGURE 39 : ACCÈS À UNE RESSOURCE ARABE VIA UNE URL LATINE/ARABE	144
FIGURE 40 : ACCÈS À UNE RESSOURCE LATINE VIA UNE URL LATINE/ARABE.....	144
FIGURE 41 : LA PARTIE REQUÊTE DANS UNE URL DYNAMIQUE	145
FIGURE 42 : FONCTIONNEMENT D'UN MODELE DNS MULTILINGUE.....	146
FIGURE 43 : LA CONVERSION DES CARACTÈRES NON-ASCII DANS LES URLs.....	148
FIGURE 44 : LE CODAGE HEXADÉCIMAL DES CARACTÈRES NON ASCII AU NIVEAU DES URLs.....	148
FIGURE 45 : CONVERSION EN %HH D'UNE CHAÎNE DE CARACTÈRES ARABES DANS UNE URL DE REQUÊTE.....	149
FIGURE 46 : CHOIX DES CODES DES CARACTÈRES SOUS MSIEXPLOREUR 4.....	151
FIGURE 47 : CHOIX DES CODES DES CARACTÈRES SOUS NETSCAPE COMMUNICATOR	152
FIGURE 48 : URL MULTILINGUE ARABE/LATIN (FICTIVE !)	153
FIGURE 49 : NUANCES ENTRE MARQUES « LRM » ET « RTL » SUR ZONES URL ET TEXTE	153
FIGURE 50 : PARAMÉTRAGE D'URLs SANS IDENTIFIANTS DES PROTOCOLES SOUS MSIE4.....	155
FIGURE 51 : LISTE DE CHOIX DE PROTOCOLES SOUS MS WORD 97	155
FIGURE 52 : STRUCTURE SIMPLIFIÉE D'UN DOCUMENT HTML	156
FIGURE 53 : EN-TÊTE D'UN MESSAGE CODE EN ISO-2022-JP TRANSMIS EN QP.....	159
FIGURE 54 : EN-TÊTE D'UN MESSAGE CODE EN MULTIPART TRANSMIS EN QP ET BASE 64.....	160
FIGURE 55 : PARAMÉTRAGE DES CHOIX UTILISATEURS SOUS OUTLOOK EXPRESS	160
FIGURE 56 : RENDU D'UN MESSAGE ARABE CODE EN QP.....	161
FIGURE 57 : SOURCE DU MESSAGE ARABE CODE EN QP	162
FIGURE 58 : LES DEUX MODES D'ENCODAGE QP ET BASE 64 DANS UN MESSAGE MULTIPART	163
FIGURE 59 : ÉCHANGE CODE EN MIME ENTRE JEUX DE CARACTÈRES IDENTIQUES	164

FIGURE 60 : ÉCHANGE CODE EN MIME ENTRE JEUX DE CARACTERES DIFFERENTS	164
FIGURE 61 : PARAMETRAGE UTILISATEUR DES MODES TEXTE ET DES TECHNIQUES DE CODAGE SOUS <i>OUTLOOK EXPRESS</i>	164
FIGURE 62 : OPTIONS DE CODAGE SOUS PEGAUS MAIL.....	165
FIGURE 63 : ÉTIQUETAGE LINGUISTIQUE SIMPLE PAR ZONE METADATA	166
FIGURE 64 : EXTRAIT DU FORMULAIRE DE DEFINITION DES ELEMENTS META PAR <i>META BUILDER</i>	167
FIGURE 65 : ETIQUETAGE LINGUISTIQUE D'UN DOCUMENT HTML ARABE PAR L'ELEMENT « META » SOUS TANGO.....	168
FIGURE 66 : MODELE DE MESSAGES ENTRE CLIENTS ET SERVEURS POUR TRAITER UNE REQUETE PAR FORMULAIRE..	170
FIGURE 67 : NEGOCIATION DE REPONSE ENTRE CLIENT ET SERVEUR A BASE DE CRITERES LINGUISTIQUES	171
FIGURE 68 : MODELE DE SERVEURS DE TRANSCODAGE INTERMEDIAIRES	174
FIGURE 69 : MODELE DE SERVEURS DE GLYPHES INTERMEDIAIRES	175
FIGURE 70 : MODELE COMBINE DE SERVEURS INTERMEDIAIRES DE TRANSCODAGE ET DE GLYPHES.....	176
FIGURE 71 : LE RENDU D'UNE CONVERSION PARTIELLE ENTRE JEUX DES CARACTERES CODES	176
FIGURE 72 : TRANSCODAGE INTEGRAL ENTRE CLIENT ET SERVEUR DE PROFIL LINGUISTIQUE IDENTIQUE	176
FIGURE 73 : LES TROIS PREMIERS NIVEAUX D'ALIGNEMENT DES TEXTES PARALLELES.....	178
FIGURE 74 : ALIGNEMENT PARALLELE ENTRE NOTES ET APPELS DE NOTES SOUS L'INTERFACE MICROSOFT WORD ...	179
FIGURE 75 : LE MECANISME DE LIENS PAR ANCRAGE MULTILINGUE	183
FIGURE 76 : MODELE STANDARD DE STRUCTURATION D'UN SERVICE HYPERTEXTE PARALLELE	185
FIGURE 77 : MODELE DE STRUCTURATION D'UN SERVICE HYPERTEXTE PARALLELE MULTIDOCUMENTS.....	185
FIGURE 78 : LES PHASES DU PROCESSUS DE TRADUCTION AUTOMATISEE	188
FIGURE 79 : CODAGE DES CARACTERES SPECIAUX ET ETENDUS PAR REFERENCES NUMERIQUES ET ALPHABETIQUE SOUS HTMLÉD.....	191
FIGURE 80 : CHOIX DE FONTE PERSONNALISEE SOUS MS-IE 40	193
FIGURE 81 : CHOIX DE FONTE PERSONNALISEE SOUS NETSCAPE COMMUNICATOR	193
FIGURE 82 : L'ASSOCIATION AUTOMATIQUE ENTRE LANGUES ET FONTES DANS LES PRODUITS UNICODE D'ACCENT SOFT.....	194
FIGURE 83 : LE TRAITEMENT DES PAGES GRAPHIQUES PAR AROBAT READER.....	195
FIGURE 84 : LA VISIONNEUSE DE POSTSCRIPT GSVIEW	195
FIGURE 85 : CARACTERES ARABES EN FONTES EXPLICITES OU PAR REFERENCES NUMERIQUES ET ALPHABETIQUES ..	196
FIGURE 86 : CARACTERES ARABES NON RECONNUS SOUS MSIEXPLORER 3.0.....	196
FIGURE 87 : PARAMETRAGE DES FONTES ARABES A LARGEUR FIXE ET PROPORTIONNELLE SOUS MSIEXPLORER 3.0	197
FIGURE 88 : RECONNAISSANCE UNIFORME DES FONTES EXPLICITES ET IMPLICITES	197
FIGURE 89 : L'ATTRIBUT "FACE" AJOUTE AUX GLYPHES ARABES REELS PAR L'ELEMENT	198
FIGURE 90 : EFFET DE L'ATTRIBUT "FACE" SUR LES GLYPHES REELS DES CARACTERES ARABES	198
FIGURE 91 : L'ATTRIBUT "FACE" AJOUTE AUX GLYPHES REELS ET AUX REFERENCES ALPHANUMERIQUES DES CARACTERES ARABES	199
FIGURE 92 : EFFET DE L'ATTRIBUT "FACE" SUR LES GLYPHES REELS ET LES REFERENCES ALPHANUMERIQUES DES CARACTERES ARABES	199
FIGURE 93 : LES DEUX TABLES D'EXTENSION DE CODES SELON ISO 8859-1 ET ISO 8859-6.....	200
FIGURE 94 : EDITION D'UN TEXTE ARABE DANS UN ENVIRONNEMENT LATIN INTEGRAL PAR REFERENCES ALPHANUMERIQUES	200
FIGURE 95 : RENDU DES REFERENCES ALPHANUMERIQUES DES CARACTERES ARABES JUSTIFIES A DROITE DANS UN ENVIRONNEMENT LATIN INTEGRAL PAR PARAMETRAGE DE FONTES PERSONNALISEES	201
FIGURE 96 : PARAMETRAGE DU RENDU D'UN TEXTE ARABE DANS UN ENVIRONNEMENT LATIN INTEGRAL PAR L'OPTION "FACE"	201
FIGURE 97 : RENDU ARABE DANS UN ENVIRONNEMENT LATIN INTEGRAL PARAMETRE PAR L'OPTION "FACE"	201
FIGURE 98 : NAVIGATEUR MULTILINGUE SUR PLATE-FORME LATINE.....	203
FIGURE 99 : EN-TETE HTTP ENRICHEE PAR LES ELEMENTS DTD ET CSS	207
FIGURE 100: TRAITEMENT LINGUISTIQUE PAR REGLES DE STYLE AU SEIN DE L'EN-TETE ET DU CORPS D'UN DOCUMENT HTML.....	211
FIGURE 101 : INDEXATION ET DESAMBIGUISATION LINGUISTIQUES PAR ATTRIBUTS « META NAME » ET « LANG » SOUS TANGO CREATOR	215
FIGURE 102 : LA DIRECTION DE BASE DEFINIE PAR L'ATTRIBUT <DIR> EN JONCTION AVEC L'ELEMENT <HTML> ...	216
FIGURE 103 : CONFORMITE DE TANGO A LA VALEUR "RTL" DE LA DIRECTIONNALITE DE BASE	218
FIGURE 104 : INADAPTATION DU NAVIGATEUR ACCENT MULTILINGUAL MOSAÏC AU PARAMETRE "RTL"	218
FIGURE 105 : DOUBLE INADEQUATION DE MSIEXPLORER LATIN 4.0 AVEC LA FONTE ARABE ET LA DIRECTIONNALITE DE BASE "RTL"	219
FIGURE 106 : ALTERNATIVE "RTL" PAR ATTRIBUT <ALIGN="RIGHT"> SOUS MSIEXPLORER 4.0.....	219
FIGURE 107 : CONFORMITE INTEGRALE DE TANGO AUX ATTRIBUTS DE BIDIRECTIONNALITE IMPLICITES ET EXPLICITES	220
FIGURE 108 : NON ACCEPTATION PAR ACCENT MULTILINGUAL MOSAÏC DES ATTRIBUTS DE DIRECTIONNALITE EXPLICITES	221

FIGURE 109 : REAJUSTEMENT DE LA BIDIRECTIONNALITE SOUS ACCENT MULTILINGUAL MOSAÏC PAR ATTRIBUTS D'ALIGNEMENT	221
FIGURE 110 : DIRECTIONNALITE "LTR" IMPLICITE DES INCISES	222
FIGURE 111 : DIRECTIONNALITE DE BASE "RTL" EXPLICITE DES INCISES	222
FIGURE 112 : LES NUANCES POLICE/CARACTERE/GLYPHE	226
FIGURE 113 : LA CORRESPONDANCE PLUSIEURS A PLUSIEURS ENTRE CARACTERES ET GLYPHES : CAS DU "LAM-ALEF"	228
FIGURE 114 : DESAMBIGUISATION PAR DIACRITISATION PARTIELLE	229
FIGURE 115 : CHOIX DE PAGES DE CODES MULTIPLES SOUS MS-DOS	231
FIGURE 116 : LES CONVERSIONS ENTRE PAGES DE CODES SOUS ARABIC MS WINDOWS	232
FIGURE 117 : LE FONCTIONNEMENT DE L'ALGORITHME D'ANALYSE CONTEXTUELLE AU NIVEAU DES PERIPHERIQUES DE SORTIE	235
FIGURE 118 : LE FONCTIONNEMENT DE L'ALGORITHME D'ANALYSE CONTEXTUELLE AU NIVEAU DE L'UC	235
FIGURE 119 : LES DIFFERENTES FORMES DES GLYPHES DE LA LETTRE « HAMZA » AU DEBUT DU MOT	236
FIGURE 120 : DES GLYPHES DE LA LETTRE « HAMZA » AU MILIEU DU MOT	236
FIGURE 121 : DES GLYPHES DE LA LETTRE « HAMZA » A LA FIN DU MOT	236
FIGURE 122 : FORMULES DE CALCUL DE L'ENTROPIE DANS L'EXEMPLE MARAYATI	238
FIGURE 123 : LA STRUCTURE COMBINATOIRE DES DIFFERENTS GLYPHES "LAM-ALEF"	238
FIGURE 124 : LE CLAVIER ARABE DANS SES DEUX MODES DE BASE ET SHIFT	238
FIGURE 125 : LE ALEF-LAM COMME SYLLABE D'UN MOT OU COMME UN ARTICLE DEFINI	240
FIGURE 126 : L'ORDRE DE TRI LATIN DANS LES DEUX ALGORITHMES DE MS-DOS ET DE MS-WORD 97	251
FIGURE 127 : RENDU DE TRI LATIN SOUS MS-ACCESS 97	252
FIGURE 128 : ANOMALIE DE L'OPTION DE RESPECT DE LA CASSE DANS LE TRI SOUS MS-WORD 6 & 97	252
FIGURE 129 : LA FONCTION SORT POUR LE TRI DE TEXTE ARABE SOUS ARABIC MS DOS	253
FIGURE 130 : LES OPTIONS DE TRI D'UN TEXTE ARABE SOUS ARABIC WORD 6 ET 97	255
FIGURE 131 : LA PRIORITE ENTRE LES OPTIONS DE TRI EN LANGUE ARABE SOUS MS WORD 97	257
FIGURE 132 : UN RENDU DE TRI SOUS CDS-ISIS ARABISE	259
FIGURE 133 : CRITERES DE TRI DE ARABIC MS WORD DANS LES DEUX MODES MAITRES	263
FIGURE 134 : LA PRIORITE ABSOLUE AUX CHIFFRES DANS LE TRI SOUS MS EXCEL	264
FIGURE 135 : MODES ET METHODES DE TRI MULTILINGUE SOUS MS DOS ET MS WINDOWS	264
FIGURE 136 : LA MULTILINGUISME DES NOUVELLES APPLICATIONS DE TRAITEMENT DE TEXTE CHEZ ACCENT SOFTWARE LTD.	268
FIGURE 137 : LE TRAITEMENT MULTILINGUE SOUS UNIVERSAL WORD	268
FIGURE 138 : LA CONVERSION ENTRE LES DEUX MODES LOGIQUE ET VISUEL EN MODE MAITRE LATIN	270
FIGURE 139 : LA CONVERSION ENTRE MODES LOGIQUE ET VISUEL EN MODE MAITRE ARABE	270
FIGURE 140 : INVERSION DE POSITION DES CHAINES OPPOSEES SELON LE MODE MAITRE ACTIF	270
FIGURE 141 : MODES DE SELECTION INTERNE (LOGIQUE) ET EXTERNE (VISUEL) EN MODE MAITRE LATIN	271
FIGURE 142 : UNICITE DIRECTIONNELLE GAD (LTR) DES CHIFFRES EN MML	271
FIGURE 143 : UNICITE DIRECTIONNELLE GAD (LTR) DES CHIFFRES EN MMA ET INVERSION SEMANTIQUE	272
FIGURE 144 : LE PARAMETRAGE DU CURSEUR EN MODES LOGIQUE OU VISUEL SOUS MS WORD 97	272
FIGURE 145 : INCISE ET BRIS EN MODES LOGIQUE ET VISUEL ERRONES	273
FIGURE 146 : INCISE ET BRIS EN MODES LOGIQUE ET VISUEL CORRECTES	274
FIGURE 147 : INCISE ARABE CORRECTE EN MODE MAITRE LATIN	274
FIGURE 148 : LE SENS DE LECTURE D'UNE INCISE ARABE EN MODE MAITRE LATIN	275
FIGURE 149 : INCISE LATINE CORRECTE EN MODE MAITRE ARABE	275
FIGURE 150 : LE SENS DE LECTURE D'UNE INCISE LATINE EN MODE MAITRE ARABE	275
FIGURE 151 : INCISE ARABE INCORRECTE EN MODE MAITRE LATIN	276
FIGURE 152 : INCISE LATINE INCORRECTE EN MODE MAITRE ARABE	276
FIGURE 153 : L'ARABISATION PAR NLSPANEL DE ARABIC MS-DOS	277
FIGURE 154 : L'EFFET DE L'IDENTIFICATION "NUMERIQUE" OU "TEXTE" DES CHIFFRES SOUS MS ACCESS	280
FIGURE 155 : PIVOTEMENT D'UNE INCISE SUR LA MEME LIGNE	281
FIGURE 156 : BRIS NORMAL D'INCISE ARABE EN MML	281
FIGURE 157 : PIVOT ANORMAL D'UNE INCISE ARABE EN MML	281
FIGURE 158 : LE PIVOTEMENT DES CHIFFRES DANS UNE SUITE ARABE	282
FIGURE 159 : LES SYSTEMES NUMERIQUES HIEROGLYPHIQUE ET HIERATIQUE EGYPTIENS	285
FIGURE 160 : LE SYSTEME NUMERIQUE ROMAIN	285
FIGURE 161 : REPRESENTATION DES VALEURS NUMERIQUES PAR BATONNETS	286
FIGURE 162 : SYSTEMES DE CALCUL 5-10 & 5-20	287
FIGURE 163 : LE SYSTEME DE COMPTE 10-60	287
FIGURE 164 : CORRESPONDANCE DE CHIFFRES ENTRE SYSTEMES ARABE DE L'OUEST ET ARABE DE L'EST	291
FIGURE 165 : EVOLUTION HISTORIQUE DES CHIFFRES ARABES ENTRE 250 A.J. ET LE 16E SIECLE	292
FIGURE 166 : LA FORME DES CHIFFRES ARABES SELON LE NOMBRE DE LEURS ANGLES	292

FIGURE 167 : L'ABAQUE	293
FIGURE 168 : MIGRATIONS DES SYSTEMES NUMERIQUES HINDOU-ARABES.....	295
FIGURE 169 : CALCULABILITE ARITHMETIQUE DE DROITE A GAUCHE.....	298
FIGURE 170 : OPPOSITION ENTRE LA TRANSCRIPTION ET LES VALEURS A POSITIONS POUR LES CHIFFRES EN MODE LATIN STANDARD	299
FIGURE 171 : ACCORD DIRECTIONNEL ENTRE TEXTE LATIN ET FACTEUR DE TRANSCRIPTION DE CHIFFRES	299
FIGURE 172 : OPPOSITION DIRECTIONNELLE ENTRE TEXTE LATIN ET CALCULABILITE DES CHIFFRES.....	299
FIGURE 173 : CONCORDANCE ENTRE LA TRANSCRIPTION ET LES VALEURS A POSITIONS POUR LES CHIFFRES EN MODE ARABE STANDARD	300
FIGURE 174 : DIRECTIONNALITE OPPOSEE ENTRE TRANSCRIPTION ET CALCULABILITE DES CHIFFRES EN MODE ARABE HYBRIDE.....	300
FIGURE 175 : POLEMIQUE DES DEUX COLONNES DE CHIFFRES AUX POIDS LES PLUS FAIBLES	301
FIGURE 176 : LE SENS DE LA TRANSCRIPTION ALPHANUMERIQUE EN MODE ARABE STANDARD.....	301
FIGURE 177 SENS DE LA TRANSCRIPTION ALPHANUMERIQUE EN MODE ARABE HYBRIDE.....	302
FIGURE 178 : MODES DE TRANSCRIPTION GRAPHIQUE DES CHIFFRES.....	302
FIGURE 179 : PRONONCIATION LITTERAIRE DES CHIFFRES.....	303
FIGURE 180 : PRONONCIATION HYBRIDE DES CHIFFRES	304
FIGURE 181 : STRUCTURE ORALE DES NOMBRES EN MODE ARABE HYBRIDE	304

CHAPITRE PREMIER

ÉTAT DE LA QUESTION

1. INTRODUCTION

1.1 UNE SOCIÉTÉ EN MUTATION : CONVERGENCE OU DIVERGENCE DE MODÈLES ?

Nul doute que parmi les euphories qui ont marqué cette fin du 20^e siècle, les turbulences et perturbations géopolitiques qui se sédentarisent dans une grande partie du monde sont témoins de la pérennité du phénomène de disconcordances culturelles, sociologiques, politiques et économiques qui ont toujours marqué le parcours de l'histoire humaine. Cette prolifération de contre-courants issus de ces disconcordances justifie d'ailleurs cette mosaïque de cultures, de systèmes, de normes et de lois qui identifient encore notre époque moderne et qui l'ont toujours accompagné durant les trois grandes révolutions agricole, industrielle et informationnelle qu'elle a connues.

Il reste cependant un fait de dire que, par comparaison avec des périodes préindustrielles du XIX^e et du début de ce XX^e siècle, la civilisation humaine a, malgré tout, déjà atteint un grade assez avancé d'universalité dans les normes et les lois qui gèrent beaucoup d'aspects de la coexistence interculturelle et de la coopération multinationale. Il apparaît en fait que par simple interaction d'opérateurs universels multiples, généralement unanimes et souvent implicites, ou même par simple logique de dépendance dictée par une verticalité de rapports stratégiques, on progresse rapidement vers ce que l'on gagne à appeler la mondialisation des modèles de sociétés, des méthodes et des systèmes. N'est ce pas là l'un des aboutissements logiques d'une « *société industrielle caractérisée par la centralisation des moyens de production, la distribution massive d'objets standardisés, la spécialisation des tâches et leur contrôle hiérarchique* »^[1] ?

Il a été un temps où cette mondialisation bifurquait en deux tendances diamétralement opposées dues à la bipolarité entre deux courants idéologiques qui ont divisé le monde en deux blocs, l'un socialiste et l'autre capitaliste. Que ce soit dans les pays à économie dirigée où les découvertes et inventions scientifiques et techniques étaient mises en exergue dans une optique propagandiste à effets immédiats ou bien dans les pays à économie libérale où la concurrence nationale et surtout internationale œuvraient pour une amélioration continue de produits et des systèmes de fabrication, le nouveau monde industriel puisait son énergie dans un antagonisme créatif et compétitif de deux modèles de systèmes économiques opposés. Depuis la fin de la première guerre mondiale, la machine de guerre venait exhauser un désir ardent d'expansion économique converti à travers les conventions et les accords multilatéraux (Sikes Piko, Yalta, Bandung...) en un mouvement de colonisation massive accompagné d'un développement parallèle d'une industrie militaire. Alors que le bloc socialiste s'est préoccupé depuis la révolution d'octobre 1917 de diriger et d'exploiter rationnellement ses découvertes scientifiques, l'occident a dû attendre les crises de 1929 et de 1940 pour décider d'une planification industrielle plus accentuée à travers les organismes de coordination nationale en matière de sciences et techniques. Alors que les inventions d'antan qui résultent souvent du simple hasard de découverte devaient attendre des années, voire même des siècles avant leur exploitation, leur application dans la vie économique devenait, à partir de 1940, presque immédiate. L'exemple très significatif de cette nouvelle structuration de la recherche scientifique est fourni par l'énergie atomique : en 1938, on

[1] Joël DE ROSNAY. Ce que va changer la révolution informationnelle. Le Monde Diplomatique, août 1996 ; p.19

réalise la première fusion du noyau de l'atome ... en 1945 ont lieu les bombardements atomiques de Hiroshima et de Nagasaki.

Le monde occidental de la post-guerre s'orientait désormais vers un modèle stéréotype de société industrielle qui s'est organisée en une pyramide de pouvoirs gérant des filières industrielles linéaires caractérisées par une hiérarchie de fonctions et pour lesquelles la recherche, le développement industriel et la multiplicité des marchés constituent les piliers d'une activité de production et distribution en masse des biens manufacturés. C'est là, en fait, un héritage économique du XIX^e siècle qui repose sur des principes tels que : «*gestion de la rareté, concentration sur la production et la distribution, spécialisation des tâches, contrôle et programmation des activités*»^[2].

Il faut attendre le début des années 1990, période marquée par la désintégration du bloc socialiste, pour que l'annonce d'une nouvelle ère commence à prendre forme sur le nouvel échiquier mondial où toute une série de transformations sous-jacentes impose partout des comportements économiques identiques sans tenir compte des particularités socioculturelles et économiques des nations^[3]. À l'instar de la doctrine d'Adam Smith, de « la révolution Keynésienne » et sa théorie générale du travail, des modèles clés en mains de systèmes économiques se sont succédés pour aboutir finalement à l'état actuel d'une économie marchande fondée sur une série de principes unanimement reconnus et souvent imposés : concurrence et compétition nationales et internationales, impôts, taxes et prélèvements sociaux.. Il s'agit actuellement de la théorie du Nouvel Ordre Mondial qui vise une reconsidération des paradigmes stabilisateurs de la société humaine dans toutes ses facettes socioculturelles, économiques et politiques selon des principes d'unification et de normalisation, de compétitivité et d'intérêt, de rivalité, de quête de suprématie et de leadership.

Cette uniformisation des standards de société relevait d'une forme de mondialisation rampante nécessaire à la promotion de la coopération internationale, laquelle sera renforcée par un ensemble de règles édictées à juste titre par l'Uruguay Round du GATT afin d'établir une libre circulation des biens et des services, et donc quelque part sur la suprématie des agents économiques les plus puissants.

Parallèlement à ce consensus « unanime » sur le bien-fondé des théories économiques entièrement développées et imposées par les pays industrialisés, l'émergence d'un déséquilibre économique régional a accentué la juxtaposition de deux blocs d'intérêts divergents : un monde industrialisé et économiquement riche et un monde en voie de développement en quête d'une stabilité économique chaque fois plus difficile à atteindre. La réalité économique mondiale se voyait de la sorte divisée en deux sous-ensembles parfois complémentaires mais souvent incompatibles : l'un bénéficiant de la suprématie économique et composé d'une minorité de pays (G7, CEE, ALENA, COMECON), l'autre engagé dans une lutte acharnée, celle des pays en voie de développement pour réduire les écarts de développement grâce à des regroupements régionaux (OEA, ASEAN, G77, CONTADORA), aux ONG et aux programmes d'aide au développement (USAID, CRDI, ENDA..), les instances internationales des Nations Unies (UNESCO, PNUD, FAO, OMS..), à la finance internationale (FMI, BM).

Marta Dosa synthétise ces différentes phases de l'évolution de la coopération internationale en quatre grandes étapes principales^[4]. Après la deuxième guerre mondiale et plus particulièrement

[2] Joël DE ROSNAY. Ce que va changer la révolution informationnelle. Op.cit.p.19

[3] Ignacio RAMONET. Québec et Mondialisation. Le Monde Diplomatique, Avril 1996 ; p.1

[4] Marta DOSA. Information Transfer as Technical Assistance for Development. Journal of the American Society for Information Science. Vol.36, n°3, 1985 ; pp.146-152.

durant les années 50 et les années 60, *la Théorie de la Modernisation* entamée par les agences spécialisées des Nations Unies et d'autres organismes internationaux, visait l'amélioration du taux de croissance économique des pays nouvellement indépendants. L'objectif était de parvenir à une amélioration du niveau de leurs conditions sociales à travers un partage équitable des richesses. Des grands projets de transfert de technologies ont été ainsi orientés vers l'implantation d'une infrastructure lourde dans des secteurs comme l'agriculture, le génie civil, la santé et les transports. Car « *la croyance occidentale selon laquelle la technologie permet de résoudre bon nombre de problèmes et de brûler les étapes était alors bien partagée* »^[5]. Tout un spectre d'information scientifique et technique accompagnait ces programmes de transfert technologique comme une nécessité absolue pour renforcer les bases économiques et industrielles des pays en voie de développement. On assistait depuis lors à la prolifération de centres et d'institutions spécialisés en information scientifique et technique, sponsorisés par des organismes internationaux essentiellement l'UNESCO.

À la fin des années 60, il était avéré que le concept de la Théorie de la Modernisation aboutissait à un déséquilibre entre le développement économique et le progrès social dû essentiellement à une marginalisation des zones rurales et à la disproportion dans le partage des ressources entre les différentes catégories sociales.

Au début des années 70, trois grands phénomènes économiques se sont manifestés dans une tentative de parvenir à un équilibre du développement régional : *politique de subvention aux besoins essentiels des pays nouvellement indépendants, Dialogue Nord-Sud et le Nouvel Ordre Économique Mondial*.

La conférence mondiale sur *l'Emploi dans le Monde*, organisée par le Bureau International du Travail en 1976, a identifié comme *Besoins Essentiels de Survie*, une alimentation adéquate, un logement et de l'habillement. Les services de bases ont été également identifiés : la garantie de l'eau potable, des services de santé, du transport public, de l'éducation, des facilités culturelles et du libre choix de travail^[6]. Une nouvelle approche a cependant marqué cette phase : accentuer les programmes de développement sur les zones rurales par des petits projets flexibles avec un taux plus élevé de participation des acteurs locaux. Une difficulté subsistait : le manque d'information sur les zones concernées et d'indicateurs de la qualité de vie des citoyens. Ceci s'est traduit immédiatement en une série de travaux prioritaires de collecte de données dirigés par des instances internationales comme l'OCDE, l'UNESCO et la Banque Mondiale afin de disposer de ressources d'information nécessaires pour la conceptualisation de programmes et de plans de développement^[7]. La conception de méthodes innovatrices de diffusion de l'information a également suscité un travail de recherche auprès des utilisateurs finaux pour optimiser l'impact des programmes d'aides au développement.

L'une des critiques formulées à l'égard de cette approche d'aide au développement, était l'état de dépendance que le nouveau système économique commençait à imposer au profit des pays riches. Cette divergence a pris forme en 1974 durant la sixième session spéciale de l'Assemblée générale des Nations unies quand « le Groupe des 77 » a entrepris la défense des pays en voie de développement dans le cadre d'une nouvelle vision d'un *Dialogue Nord-Sud*. Cette orientation portait du contrôle économique et politique qu'exerçaient les pays industrialisés et les disparités

[5] Yvonne MIGNOT-LEFEBVRE. Technologies de communication et d'information : une nouvelle donne internationale ? Revue Tiers Monde, t.XXX, n°138, avril-juin 1994. p.249

[6] International Labor Organization. Employment, Growth and Basic Human Needs : A One-World Problem. Genève : BIT, 1976

[7] N. HICKS & P. STREETEN Indicators of Development : the Search for a basic needs yardstick. World Development. N°7, 1979 ; pp. 567-580.

observées en terme de recherche scientifique et de développement technologique pour proposer une reconsidération de ces deux facteurs essentiels pour l'amélioration des conditions économiques et sociales dans les pays en voie de développement.

La proposition d'un *Nouvel Ordre Économique Mondial* a surgi des cendres du Dialogue Nord-Sud devenu rapidement obsolète. Depuis 1974, le programme d'action de l'Assemblée générale des Nations Unies pour un Nouvel Ordre Économique Mondial portait dans ses replis de restructuration, des recommandations aux pays en voie de développement afin de "*promouvoir et d'établir des instruments efficaces de coopération dans les domaines de l'industrie, de la science et de la technologie, des transports, de la navigation et des moyens de communication de masse*"^[8]. L'un des objectifs de la conférence de Buenos Aires (1978), soit le soutien des pays en voie de développement afin de parvenir à une indépendance nationale et collective^[9], était confié au PNUD par l'Assemblée générale des Nations Unies afin de mettre en œuvre une stratégie compensatrice de l'échec des modèles de développement proposés par le Nord. Une remise en question de la distribution internationale de l'information scientifique et technique, des modalités du transfert de technologies et du contrôle des réseaux de communication^[10] était alors indispensables.

En effet, au lieu de rester en position catalysante par rapport aux processus de croissance des pays, aux projets de restructuration économique et aux programmes d'aide au développement, l'information était désormais devenue l'objectif même des négociations économiques et commerciales donnant naissance à des secteurs d'activité stratégique comme l'industrie de l'information et l'économie de l'information. À partir des années 80, le phénomène du transfert de technologies avait consolidé concrètement cette nouvelle donne grâce à la microminiaturisation des outils électroniques. Avec la vulgarisation de l'outil informatique et sa libération des enceintes des laboratoires et des services hautement qualifiés, les planificateurs ont pu se rendre compte directement et apprécier le rôle de l'information dans tous les projets qu'ils entreprenaient. L'usage généralisé des réseaux de télécommunications a donné une autre empreinte d'efficacité à l'information scientifique et technique car il constitue aujourd'hui l'image, le symbole même d'une société universelle et intercommunicante. Ce sont là en fait les prémisses d'une nouvelle révolution, la révolution informationnelle que Toffler et Surprenant ont identifié comme une troisième vague de réformes d'envergure universelle après les deux révolutions agricole et industrielle^{[11], [12]}.

1.2 LA REVOLUTION DE L'IMMATÉRIEL : UNE NOUVELLE DONNE SOCIO-ECONOMIQUE

Suite à ce constat introductif relatif à la scène économique internationale et aux nouvelles exigences de son évolution, nous allons aborder certains aspects plus spécifiques du nouveau contexte informationnel mondial, clarifier notamment les rapports intrinsèques et les effets de causalité entre ses différentes phases, tenter de comprendre la logique des liens qui gagnent en rigueur entre des concepts émergents tels : Société d'Information ; Information Scientifique et Technique ; Communication ; Nouvelles Technologies de l'Information ; Interfaces et langages de communication. L'objectif demeure la présentation d'un contexte global d'analyse qui

[8] Ruth SELIGMAN. The Buenos Aires plan of action for promoting and implementing technical cooperation among developing countries : ten years after. Cooperation south. Décembre 1977 ; p. 11

[9] Technical cooperation conference adopts plan of action : self-reliance stressed. UN Chronicle, octobre 1978 ; pp. 43-46

[10] International Commission for the Study of Communication Problems (MacBride Commission). Many Voices. One World. Paris : Unesco, 1980.

[11] Ochai ADAKOLE. The emerging information society. International Library Review, N° 16, 1984 ; pp.367-372

[12] Tom SURPRENANT. Future libraries : the electronic environment. Wilson Library Bulletin, Janvier 1982.

servirait de plate-forme de départ pour cibler un aspect plus pointu : celui des interfaces Homme-Machine dans le processus de communication de l'Information Scientifique et Technique, et ce, dans un contexte linguistiquement hétérogène et complexe comme celui des pays de langue arabe. Ceci nous aidera également à trouver les arguments nécessaires et les justificatifs adéquats à des emprunts épistémologiques à l'égard de disciplines annexes comme l'économie, l'informatique, la sociologie ou la linguistique. Ce sera en quelque sorte un ensemble de considérations et de constatations qui trouveront leurs prolongements dans des descriptions techniques d'artefacts choisis comme supports d'analyse et de démonstration.

1.2.1 LA SOCIÉTÉ DE L'INFORMATION : MYTHE OU RÉALITÉ ?

Nous partirons du constat effectué de l'état de prépondérance de l'information dans la société contemporaine pour définir les origines de ce qui est communément appelée "*une société de l'information*", ses critères, ses implications et ses limites.

1.2.1.1 Origines :

Comme l'annonçait Le Coadic en 1994, la Société de l'Information est l'aboutissement du développement rapide de la consommation des produits informationnels engendré par l'émergence de ces produits "*dans la sphère de la production et de l'échange marchand donnant naissance à ce qu'on appelle les industries de l'information, le marché de l'information, avec leur cortège de bien informationnels, de service informationnels, de produits informationnels, tous plus ou moins assistés par ordinateurs*"^[13]. Ce nouveau phénomène justifie, comme l'ont démontré Nora et Minc^[14], le recours à l'arme informationnelle et à l'arsenal technologique pour créer un nouveau contexte économique mondial et une nouvelle forme de croissance basée sur les nouveaux services et produits d'information. La qualité de la vie et les projets de réformes sociales et de développement économique dépendent désormais des ressources d'information et de leurs applications. Les standards de vie, ainsi que les modèles de travail et de divertissement, la nature des marchés, les systèmes éducatifs et la recherche sont tous influencés par les progrès de l'information et de la connaissance qui en découlent^[15]. Dordick considère que ce stade d'évolution résulte de l'enclenchement d'un mécanisme sociétal qui progresse graduellement à travers les étapes clés : la prise de conscience de la valeur de l'information, considère le développement d'un milieu favorable au traitement de l'information, le développement d'un secteur économique spécifique^[16]. Dans la perception d'Ochai, le vrai parcours passe obligatoirement par un état de société consciente des enjeux de l'information, puis par un état de société ayant une culture de l'information pour aboutir en fin de compte à un état de société d'information dans laquelle l'accès instantané aux ressources deviendrait une affaire ordinaire^[17].

[13] Yves-François LE COADIC. La science de l'information. Paris : PUF, 1994.-(Que Sais-je?, 2873).-p.3

[14] S. NORA, A. MINC. L'informatisation de la société. Paris : La Documentation Française, 1978.

[15] William J. MARTIN. The information society : idea or entity. Aslib Proceedings. Vol.40, n°11-12, novembre 1988 ; pp.303-309.

[16] Herbert DORDICK. Information society indicators : description, measurement, prediction. In : Information societies : comparing the Japanese and American experiences. Seattle : International Communication Center, 1978.

[17] Adakole OCHAI. The emerging information society. International library review. N°16, 1984 ; pp.367-372.

1.2.1.2 Critères :

Face à ces controverses dans la caractérisation d'un phénomène, certains auteurs esquissent un modèle de société qui voit les aspects technologiques s'enchevêtrer avec les critères sociaux pour envisager une convergence possible des politiques et des technologies d'information et de communication vers des secteurs de développement économique et social (i.e. le rapport Nora Minc). Dans ce sens, Martin a établi une grille d'évaluation approximative des critères et aspects permettant d'identifier un statut équilibré d'une société d'information modèle^[18].

Critères technologiques	● Une Technologie de l'Information comme force motrice, ● Généralisation des Technologies de l'Information dans les bureaux, les usines, l'enseignement et les foyers.
Critères sociaux	● L'information comme promoteur de la qualité de vie, ● Conscience générale de l'importance de l'information, ● Accès des utilisateurs finaux à une information spécialisée.
Critères économiques	● L'information comme un facteur économique clé : en tant que ressource, service, commodité, source de valeur ajoutée et d'emploi.
Critères politiques	● Une liberté d'information capable d'accentuer la participation au processus politique.
Critères culturels	● La reconnaissance de la valeur culturelle de l'information par la promotion de sa contribution au développement de la communauté et de l'individu.

Tableau 1 : Les critères de développement d'une société d'information

Source : William J. MARTIN. *The information society : idea or entity*. p.305

À ces critères, il rajoute d'autres conditions concernant le nombre de terminaux et d'ordinateurs de tout genre^[19], le taux des flux transfrontières de données, les statistiques relatives à l'industrie de l'information, la part de l'information dans le Produit National Brut et la masse de travail, le degré de recours aux médias par les personnes intervenant dans les secteurs de l'éducation et de la formation continue.

1.2.1.3 Implications :

Entre le pessimisme de Robbins^[20], argumenté par une subordination plus profonde des besoins et des valeurs sociales aux implications des technologies et des besoins de marché, et l'optimisme de Masuda^[21], issu d'une certitude de transformation radicale de la société sur des modèles structurés et ordonnés, notre position ne saura échapper à la considération d'un certain nombre d'éléments régulateurs ou perturbateurs d'ordre technique, commercial et humain qui touchent l'individu et la communauté à la fois.

[18] Martin J. WILLIAM. *The information society : idea or entity*. Op. Cit.

[19] Voir aussi Eugene GARFIELD. 2001 : an information society ? *Journal of Information Science*. N°1, 1979 ; p.210.

[20] K. ROBBINS. *New technology : the political economy of general Ludd*. In : *Information technology : impact of the way of life*, vol.2. Dublin : National Board for science and technology, 1981.

[21] Y.C. MASUDA. In : *The Information technology revolution*. Ed. by Tom Forrester, Oxford, Blackwell, 1985 ; pp.620-634.

Sur un plan individuel, des polémiques profondes relatives à la liberté d'information et à la vie privée des gens marquent encore les controverses, dans des sociétés démocratiques, entre le droit inaliénable à l'information à caractère public et les démarches de protectionnisme appliquées pour des raisons de sécurité ou de droit privé. Nul n'ignore désormais qu'à part sa généralisation en tant que ressource consommable, sa désintermédiation grandissante dans des processus communicationnels, l'information est susceptible d'agir directement sur la rationalisation des comportements décisionnels et productifs des individus.

L'aliénation constitue par contre, un revers négatif qui affecte certaines catégories d'individus. Par aliénation nous entendons l'isolement progressif des personnes non impliquées dans cette révolution technologique massive qui rend leur domination probablement irréversible. Ne parle-t-on pas d'analphabétisme informatique pour désigner ceux qui au contraire des « happy fews » sont dépourvus de toute connaissance technologique.

Abordée dans sa dimension nationale, la société de l'information à engendré, comme expliqué auparavant, une nouvelle restructuration de l'activité économique basée sur de nouvelles formes de production et de concurrence commerciale essentiellement rattachées à des produits d'information. Il est vrai que la société de l'information, avec toute l'armada infrastructurelle qu'elle engendre a contribué au développement du standard de vie de l'humanité entière (progrès médicaux, abondance agricole, innovations industrielles, échanges culturels...). Sauf que, comme le revendique Lempen, si « *le développement des nouvelles technologies a, certes, stimulé l'activité, accru la productivité dans certains secteurs, il a aussi été accompagné d'un fort taux de chômage... Des inconvénients sont apparus : civilisation de la panne, a-t-on écrit, stress psychologique et physique lié à l'utilisation intensive des ordinateurs, accroissement de la monotonie du travail avec un degré d'automatisation plus poussé.. dysfonctionnements techniques sources de conflits dans la production et la gestion* »^[22].

Autres implications liées à cette transformation du système économique mondial :

- La mise en cause des souverainetés nationales par les interactions et les échanges effectués à travers les entreprises multinationales. Du fait de la nécessité incontournable de coopération dans un environnement communicationnel universel, de l'arsenal électronique déployé par les multinationales, certains pays voient leurs activités économiques contrôlées depuis l'étranger.
- La dépendance plus grande envers les ressources électroniques d'information et les systèmes électroniques de communication. Cette dépendance, en cas de chute ou de défaillance de ces systèmes, engendre des pertes irrécupérables de données stratégiques qui peuvent conduire à des perturbations économiques et sociales importantes. La perte d'un fichier central d'une banque ou la destruction d'un registre civil électronique, peut causer, en cas de négligence de duplication ou de défaillance de sauvegarde, la perte définitive de données stratégiques.
- Le développement d'une nouvelle forme de « colonialisme » électronique matérialisé par la domination du monde occidental sur le marché de l'information et le problème compliqué du déséquilibre des flux transfrontières de données. Les pays en voie de développement sont obligés d'acquiescer chèrement des données scientifiques qui répondent aux intérêts économiques et politiques des pays industrialisés plutôt qu'aux leurs. Selon Jean Devèze,

[22] Blaise LEMPEN. La prothèse informatique. Nouvelles de la science et des technologies, vol.11, n°1, 1993 ; pp. 75-80.

« les puissances du « G7 » poussent parfois même leur « générosité » jusqu'à offrir gratuitement l'accès à certaines banques de données aux PMA »^[23].

1.2.2 SCIENTIFICITÉ ET TECHNICITÉ DE L'INFORMATION

Si l'usage du concept d'Information Scientifique et Technique remonte, à en croire Dahmane^[24], à 1952 date de création du Centre National de Documentation Soviétique VINITI (Vsesojuznyj Naucnoji Techniceskoj Informacii), les définitions qu'on en donne depuis restent encore imprécises.

Dans son analyse, Dahmane expose deux perspectives explicatives du concept d'information scientifique et technique. Une première approche, se basant sur la structure syntaxique du concept (substantif qualifié par un adjectif), définit l'IST comme étant une « *connaissance référant à tous les éléments cristallisant le contenu dénotatif/connotatif relatif à la science, la technique et la technologie* »^[25]. De la sorte, il en fait une substance soumise à deux systèmes techno-sociaux, l'un dit primaire concernant l'édition scientifique et l'édition de littérature grise ; l'autre dit secondaire, touchant les opérations de collecte, traitement et diffusion de l'information nécessaire à l'utilisateur.

La seconde approche part du néologisme du concept d'information scientifique et technique pour donner plusieurs sens au soubassement connotatif observé. Un premier sens donne à l'information scientifique et technique une dimension de spécialisation qui vise un public sélectif à travers des systèmes de communication spécifiques. Un second sens, établit une connotation entre IST et activités socio-économiques stratégiques donnant naissance à un concept nouveau d'ISTE (Information Scientifique et Technique Économique). La troisième connotation déduite stipule que l'IST est en fait le produit d'un nouveau corps professionnel, en l'occurrence la documentation qui, pour ce faire, fait recours à l'outil informatique.

C'est cette dernière interprétation que Dahmane supporte malgré cette équivoque sémantique de l'IST qu'il décrit. Il considère en effet, que l'IST demeure à la fois un produit informationnel et une activité propre aux systèmes documentaires relevant du domaine des sciences de l'information. C'est dans cette définition que se justifient les orientations actuelles des Sciences de l'Information et de la Communication (SIC) pour étudier les comportements des utilisateurs et leurs pratiques informationnelles, les techniques de traitement, de stockage, de diffusion et de restitution de l'information par les moyens informatiques.

Il est également intéressant, dans ce sens, d'observer l'analyse de Jean Devèze qui, en partant d'un antagonisme de conception entre la scientificité d'Auguste Comte et le structuralisme de Noam Chomsky, se pose des questions d'ordre conceptuel relatives aux statuts et aux conditions favorables ou défavorables pour qualifier ou même appréhender, fut-il de façon intuitive, une information quelconque dans une dimension scientifique et technique.

Le découpage du champ de la connaissance a été pour longtemps soumis à la très séculaire « classification des sciences », effectuée par Auguste Comte il y a déjà un siècle et demi. Jugé totalement arbitraire et démuné de vérité épistémologique absolue, ce découpage idéologique est

[23] Jean DEVEZE. L'information scientifique et technique : produit scientifique ou discours sur la science ? Communication et langages. N°90, 1991 ; pp.95-103.

[24] Majid DAHMANE. Information scientifique et technique (IST). Communication et langages. N°90, 1991 ; pp.104-105.

[25] Majid DAHMANE. Op. Cit. p.104

mis en cause par Devèze, qui relativise sa portée face à une « *scientification galopante* » du savoir humain. À part l'émergence de « nouvelles sciences » du genre « science du langage », « sciences de l'éducation », « sciences des religions », « sciences de la société », de plus anciennes disciplines jusque-là jugées « culturelles » se sont faites progressivement scientifier pour donner naissance à des disciplines du genre « sciences juridiques », « sciences administratives », et même paradoxalement « sciences littéraires »!!!. Les techniques ont également connu depuis la révolution industrielle, une évolution « inflationniste » de leur portée conceptuelle donnant naissance à des pseudo-disciplines techniques du genre « techniques sportives », « techniques du corps », « techniques de la traduction ».

Ces métamorphoses peuvent donner lieu à des interrogations sur les raisons de l'amenuisement, voire de la mise à l'écart d'un type d'information concernant certains secteurs de la connaissance humaine. Ce mécanisme d'octroi de priorité scientifique est expliqué par Devèze par le phénomène d'utilité affectée à une information quelconque. Ce mécanisme se base sur deux modes d'évaluation : une évaluation par le succès d'une information dans la création d'un contexte compétitif ayant pour conséquence « *la régulation de la création et de la diffusion de l'IST par le marché* » ; et une évaluation par l'imposition d'un type d'information présélectionnée et monopolisée par une entité influente. À ces deux modes d'évaluation par l'utilité s'ajoute, dans une optique de valeur marchande de l'information, un mode d'évaluation en fonction de la rentabilité de l'information.

Ainsi, « *le développement de l'IST n'a pu se faire et ne continuera à se faire que par une succession de processus de sélection, de plus en plus complexes et diversifiés destinés à procurer à l'utilisateur une information réputée "utile"* ». Devèze cite Roqueplo^[26] qui remarque que « *le discours scientifique est une construction qui ne s'impose nullement comme manifestation de la réalité, mais qui se propose comme un immense échafaudage.. certes investi d'une intention de vérité* »^[27] mais dont les différentes strates reflètent, à des niveaux constamment mouvementés, des degrés d'importance toujours mis aux enchères en présence d'opérantes économiques et politiques déterminantes.

Rouban rejoint cette approche en confirmant que « *la culture scientifique et technique s'est moulée à chaque époque sur un champ scientifique dominant : la physique nucléaire, l'informatique, la biologie ... se sont ainsi succédées depuis une cinquantaine d'années comme sources de représentations dominantes sur ce qu'était la science et les conduites légitimes qui s'y rapportait. La culture scientifique et technique connaît donc des forces et des faiblesses des disciplines qui la sous-tendent* »^[28].

C'est également dans cette optique que Habermas explique de son côté le rôle des idéologies dominantes qui imprègnent tout le discours politique et économiques des États, dans la consécration d'un type particulier d'information « scientifique et technique »^[29]. Le changement volontariste et participatif (direct ou indirect) du public dans le choix des options scientifiques, comme initialement appréhendé depuis les années 70, a fini par céder devant les exigences de l'industrialisation et de l'économie de marché. Le référent économique semble l'avoir définitivement emporté sur toutes les considérations agissant antérieurement dans la définition des politiques de la science et de la technologie. De la même façon, le projet de l'autogestion de

[26] P. ROQUEPLO. Penser la technique. Paris : le Seuil, 1983 ; p.43.

[27] Jean DEVEZE. L'information scientifique et technique : produit scientifique ou discours sur la science ? Op. cit. ; p.100

[28] Luc ROUBAN. Commentaire général : quelle politique pour l'information scientifique et technologique ? Nouvelles de la science et des technologies. Vol.1, n°1, 1993 ; pp.81-87.

[29] J. HABERMAS. La technique et la science comme « idéologie ». Paris : Gallimard, 1973.

l'activité de recherche a évolué de telle sorte que nous assistons à une réconciliation massive des chercheurs avec les entreprises dans le cadre d'un mouvement massif de privatisation.

Pour ne rester que dans l'afférent aux domaines d'application de l'IST, la littérature est également confuse à l'égard de deux pratiques informationnelles fort différentes. D'une part, une activité purement professionnelle couvrant les activités dites scientifiques de recherche, de gestion de l'entreprise, de l'administration *« conçues suivant les règles de chaque discipline et structurant le champ professionnel par l'effet-retour sans que la nature « scientifique » de l'information ait besoin d'être évoquée »*. D'autre part, une activité aussi professionnelle et technique implique *« la seule diffusion d'une information homogénéisée à destination d'un public récepteur passif »*^[30]. En d'autres termes, le domaine de l'IST laisse envisager deux univers d'usage normalement complémentaires mais qui s'affrontent déjà dans la sphère de l'activité communicationnelle : l'univers de l'information grand public qui est en réalité une « information SUR la science » retraçant les résultats obtenus par les laboratoires ou les entreprises et le monde plus restreint de la recherche scientifique et technique qui conduit une activité d'« information DE la science » visant la construction des faits scientifiques et techniques dans leur réalités parfois cachées derrière des considérations tenant à la validité économique des projets ou à leur fiabilité intrinsèques.

Cette dynamisation conceptuelle et épistémologique de l'IST a fait surgir le phénomène nouveau de « la culture informatique » rapprochant davantage les écarts entre une vision purement technicienne de la technologie et une approche sociale étudiant les techniques comme un ensemble socioculturel. Flichy argumente cette approche en précisant qu'« *en instituant une séparation totale entre la conception de l'objet technique et sa diffusion, on oublie que la conception est liée à une certaine représentation des usages et qu'à l'inverse l'appropriation par les utilisateurs n'est pas sans influence sur l'évolution du système technique* »^[31].

En fait, la littérature aborde ce phénomène de « culture informatique » dans une utopie économique basée sur un discours promotionnel des effets positifs de l'outil informatique sur la société. C'est la thèse d'un mouvement d'idées qui résume cette utopie en deux points essentiels : *« la société doit relever le défi informatique si elle veut sortir de la crise ; l'informatique est la solution aux problèmes de l'humanité »*^[32].

La quasi unanimité de cette vague d'informatisation, qui d'ailleurs succède à un état de rejet de la « mégamachine » dû à une vision « idyllique » de « l'informatique et liberté », des dégâts du progrès technologique ... explique le degré très élevé de popularité de la microinformatique justifié par son efficacité, le prestige qu'elle confère, ses aspects ludiques ainsi que les besoins structurels et conjoncturels auxquels elle répondait (crises organisationnelle et économique). Le discours idéologique dominant impose l'ordinateur comme la solution unique aux problèmes de la société industrielle. C'est la reprise des théories de Harold Innis et Marshall Mac Luhan^[33] qui prévoient que chaque moyen de communication structure les modes de connaissance et plus largement les formes d'organisation sociale. Selon ce même discours, se passer des nouvelles technologies, c'est régresser à l'âge de pierre, perdre la compétition internationale, tomber dans le sous-développement. En ce qui concerne l'individu, l'accoutumance à l'outil informatique

[30] Luc ROUBAN. Commentaire général : quelle politique pour l'information scientifique et technologique ? Op. cit. ; p.83

[31] Patrice FLICHY. La question de la technique dans les recherches sur la communication. Réseaux Communication Technologie Société. N°50, 1991 ; pp.51-62.

[32] Blaise LEMPEN. La prothèse informatique. Op.cit. ; p.78.

[33] Les deux appartenaient à l'École de Chicago pour laquelle la communication n'a pas eu comme simple objectif la transmission des messages, mais aussi comme processus symbolique qui permet à une culture de se construire et se maintenir.

passer par tous les contextes socioprofessionnels, depuis l'école jusqu'à l'entreprise. Tout attachement ou abnégation de s'y insérer engendre un phénomène d'aliénation et de marginalisation que J. De Noblet, l'un des partisans du mouvement en faveur de la culture technique décrivait en ces termes :

« Il semble évident que celui qui manque de cette culture vit dans l'ignorance de son propre milieu. Il se trouve doublement aliéné : d'une part, il ne maîtrise pas son propre environnement et d'autre part, cette absence de maîtrise le place socialement dans une dépendance permanente par rapport aux organisations et aux individus qui possèdent les compétences qui lui manquent. Il ne peut contrôler la manière dont ceux-ci exercent ces compétences, en intervenant sur sa propre existence » ^[34].

Il s'agit moins en définitive de penser l'informatisation de la société toute entière comme un processus d'adaptation technique des postes de travail par exemple que de lui reconnaître un statut de discipline scientifique enseignée selon les règles académiques ou en qualité d'une discipline dont les concepts fondamentaux devraient être appréhendés préalablement à son usage.

Les quatre générations que l'informatique a connues montrent une évolution constante tant du logiciel que du matériel vers la superposition de couches applicatives entre le noyau de la machine et son côté frontal entraînant une modularité extrême dans les composantes du système informatique et portant une plus grande facilité d'usage, pouvant faire l'économie notamment de l'apprentissage algorithmique et de la programmation. La « scientificité » de l'informatique y perd son rapport à l'information et à la communication. Avec l'avènement des interfaces graphiques et des applications multimédias l'outil informatique élargit et son public et son univers d'applications. Il a réussi la « prothèse » ^[35] apparaissant comme déterminant pour l'industrialisation de la société. Il fait désormais l'objet d'un discours idéologique très favorable.

1.2.3 INFORMATION ET NOUVELLES TECHNOLOGIES ASSOCIÉES

C'est à partir de ce point de fusion entre information et technologies associées, mieux identifiée par le concept de Nouvelles Technologies de l'Information (NTI), que nous allons interroger sur les véritables paradigmes de leur internationalisation et les enjeux stratégiques attachés à la cristallisation des caractères socioculturels locaux qui résistent difficilement à l'effet « d'effacement » engendré par une modélisation universelle unique ^[36]. « Transfert de technologies », « Flux Transfrontières de Données », s'inscrivent aujourd'hui dans la mondialisation de sociétés communicantes au risque de voir surgir à nouveau les « vieilles » querelles relatives à la dépendance technologique et culturelle, à l'inégale répartition des ressources d'information et des moyens de son exploitation entre les pays industrialisés et les pays en voie de développement.

L'information n'a pas le même statut, le même seuil d'universalité et d'acceptation. Dans sa portée universelle, elle est enrobée dans une parure culturelle, spirituelle et parfois mystique, elle fait souvent face à une prise de position de la part du réceptacle pour être porteuse d'un contenu,

[34] J. DE NOBLET. La technique dans la culture. Esprit. Octobre 1982.

[35] Blaise LEMPEN. La prothèse informatique. Op. cit. ; p.78.

[36] Yves-François LE COADIC. La science de l'information. Op. Cit.

d'un sens, d'un symbole, contrairement au support matériel qui est souvent pris dans sa dimension « *médiatrice exclusive* »^[37] catalysante, impartiale et maîtrisable.

L'information et sa technologie d'usage suscite inmanquablement une dialectique entre le culturel et le technologique, le spirituel et le pragmatique, l'abstrait et le palpable.

1.2.4 INFORMATION ET COMMUNICATION : LE DUO INSEPARABLE

Un élément important, entièrement impliqué dans le maillage techno-informationnel est constitué par le mode de communication. La communication de l'information constitue, dans la sphère des sciences sociales^[38], une discipline à part entière qui a toujours soulevé des polémiques chaudes tant au niveau de sa représentation conceptuelle qu'au niveau des domaines de son application. Deux écoles illustres, en l'occurrence l'École empirique de Chicago, constituée autour des théories de Lazarsfeld et l'École critique formée d'anciens disciples de l'École de Francfort (Marcuse, Adorno, Horkheimer, Fromm), ont conduit pour longtemps la saga des confrontations des recherches dans le domaine de la communication. « *Si la première fut caractérisée, dans ses grandes lignes, par la méthode quantitative, le fonctionnement et le positivisme, les adeptes de l'école critique privilégiaient, dans l'analyse, le contexte social dans lequel se déroule l'acte de communication* »^[39].

Ainsi fut toujours posée la question de la sociabilité de l'acte de communication qui influe et obéit désormais aux exigences imposées par l'intrinsèque liaison qui unit ses deux « congénères » information et médias. Avec l'avènement des machines à communiquer, cette complicité ne fut que davantage confirmée grâce, entre autres, aux recherches de Mac Luhan qui accorderont une place centrale à ces machines dans la détermination des différents aspects de la société. L'objet de la communication, contrairement aux autres disciplines des sciences sociales qui suivent « *les règles bien connues régissant les agents innovateurs* »^[40], jouit de certaines particularités fondamentales. D'une part il provoque des effets de réalité par son intégration aux circuits économiques et aux enjeux marchands ; de l'autre il constitue un terrain favorable d'observation des effets des sciences sociales grâce à son usage généralisé aux domaines sociaux les plus divers.

Abordé dans ses particularités contemporaines, le domaine de la communication laisse envisager deux univers d'usage complémentaires qui revendiquent explicitement l'élaboration d'un savoir théorique : celui de la communauté académique et de la recherche scientifique et technique qui affronte un univers plus éclaté des communautés professionnelles des mass-media, des publicitaires et des consultants. Tous deux, consommateurs par excellence de ce dualisme enchanteur entre l'information et les NTI, bien qu'à des fins parfois entrelacées mais principalement différentes, bifurquent en aval du processus communicationnel pour adopter, selon Neveu et Rieffel^[41] deux modèles de communication qui diffèrent en fonction des

[37] Lucien SFEZ. La communication. Paris : PUF, 1994. (Que sais-je).127 p.

[38] Pourtant, Judith Lazar, dans son Que Sais-je ? n°2634 de mai 1992, cite Wilbur Schramm avoir écrit dans les colonnes de *The Science of Human Communication* que « la communication n'est pas une discipline académique, mais plutôt une discipline de carrefour où nombreux sont ceux qui y passent, mais peu qui s'y arrêtent ». Ceci est d'autant plus vrai si l'on considère que les pères fondateurs de la science de la communication ne sont autres que le politologue Lasswell, le psychosociologue Kurt Lewin, le psychologue expérimental Carl Hovland et l'illustre sociologue Paul Félix Lazarsfeld,.

[39] Judith LAZAR. La science de la communication. Paris: PUF, 1992. (Que sais-je, 2634). p.18.

[40] Erik NEVEU, Rémy RIEFFEL. Les effets de réalité des sciences de la communication. Réseaux Communication Technologie Société. N°50, 1991 ; pp. 13-39.

[41] Erik NEVEU, RIEFFEL Rémy. Les effets de réalité des sciences de la communication. Op.cit. ; p.23

procédures de médiation utilisée ainsi que de la nature du public desservi. Le premier, identifié comme « *modèle mécaniste* », s'inspire des deux modèles de communication de Shannon et Weaver (1949) ^[42] et de Riley et Riley (1956) ^[43], pour suggérer une diffusion linéaire des connaissances de la part des scientifiques « *à un micromilieu de spécialistes et de médiateurs qui à leur tour les divulguent au grand public* ». Le deuxième modèle dit « *modèle de la concurrence* » plus proche du « *modèle de dépendance* » développé par Ball-Rockeach et Defleur (1976) et décrit par Lazar ^[44], fait appel à des systèmes de communication très impliqués dans des conflits opposant les chercheurs et les universitaires d'une part, et les professionnels des médias, de la publicité et de la consultance d'autre part.

D'un point de vue opérationnel, toute information à vocation générale ou spécialisée, n'a aucune valeur réelle si elle n'est pas communiquée dans les règles que le contexte techno-culturel dicte. Ces règles constituent au fait une série de paramètres qui reflètent, en terme d'efficacité et de rentabilité de la communication, l'état des rapport d'usage entre information et technologie. Cet état doit suggérer, les cas de bons rapports, des résultats qui garantissent, au niveau de l'information, des critères d'exhaustivité, de pertinence, de fiabilité et d'actualité, et au niveau de la technologie, des critères de rapidité, de pérennité, de réédition, de reformatage, de portabilité et d'échange. Selon M.G. Lundu, personne ne pourrait remettre en question le fait que l'information est un phénomène social, dynamique, universel, multidimensionnel et plein de valeurs ^[45]. Seulement, l'information bien qu'elle soit universelle, et multidimensionnelle de nature, ne peut avoir de sens sans être associée à un autre concept aussi universel et multidimensionnel comme celui de la communication. Luhmann précise que « *la communication ... doit dédoubler l'information : d'un coté, elle doit la laisser dehors ; de l'autre, elle doit l'utiliser pour communiquer et lui donner, en conséquence, une seconde forme appropriée, par exemple une forme langagière* » ^[46].

Devèze définit la communication comme étant « *une interaction communicationnelle des symboles. Le choix de ces symboles n'a rien d'innocent puisqu'ils sont arbitrairement définis par les émetteurs et, en amont par ceux qui élaborent les informations, les construisent, les ordonnent, les hiérarchisent, les mettent en forme* » ^[47]. Cette interaction est porteuse de tout un système de représentations de l'activité scientifique et de l'activité informationnelle destinées à être transmise au destinataire. Elle porte impérativement en elle, pour une part significative, la marque, le point de vue, les préoccupations et l'ensemble des valeurs culturelles, économiques, politiques et idéologiques dominantes des émetteurs qui ont « *tout seul, le pouvoir de garantir l'objectivité* » ^[48]. Lazar rajoute qu'en aval de ce processus, les mêmes valeurs interviennent de façon non négligeable de la part du récepteur dans le décodage de cette représentation. « *Ainsi le simple décodage d'un message n'est pas suffisant pour un individu n'appartenant pas à la même culture que l'encodeur ou l'émetteur* » ^[49] car comme le confirme Flichy « *les mêmes idées seront exprimées de façon différente dans deux subcultures distinctes* » ^[50].

[42] Judith LAZAR. La science de la communication. Op.cit. ; p.102.

[43] Judith LAZAR. La science de la communication. Op.cit. ; p.113.

[44] Judith LAZAR. La science de la communication. Op.cit. ; p.120.

[45] M.G.LUNDU The justification of establishing and maintaining a nation-wide library service ; with special reference to the library and the process of national development in Zambia. Thèse de Ph.D., Université de Sheffield, 1982.

[46] Nikolas LUHMANN. Communication et action. Réseaux Communication Technologie Société. N°50, 1991 ; pp. 131-156.

[47] Jean DEVEZE. L'information scientifique et technique : produit scientifique ou discours sur place ? Op.cit ; p.102.

[48] Lucien SFEZ. La communication. Paris : PUF, 1994. (Que sais-je?), p.14.

[49] Judith LAZAR. La science de la communication. Op.cit. ; p.104.

[50] Patrice FLICHY. La question de la technique dans les recherches sur la communication. Op.cit. ; p.58.

Sans vouloir entrer pour le moment dans des considérations techniques ni linguistiques ou cognitives pour évoquer les problèmes d'un tel codage, (nous le ferons de manière détaillée dans les parties qui suivent) nous précisons uniquement qu'il produit, en fonction du degré de son opacité, une différenciation des processus de communication et de compréhension de l'information. Le phénomène de la compréhension, tributaire en général des paramètres de la source de communication, dépend aussi d'une stratification chez le récepteur des capacités de déchiffrement et des taux de réception influencées à leur tour par le mécanisme d'intermédiation séparant l'émetteur et le récepteur de l'information. En d'autres termes, c'est au niveau des interfaces logiques ou techniques que se situent désormais les vrais enjeux de la communicabilité ou de l'incommunicabilité de l'information. Avec les nouveaux systèmes de communication basés sur les nouvelles technologies de l'information et leur nature interactive, l'émetteur et le récepteur se libèrent du modèle de communication à sens linéaire de Shannon et Weaver (Fig.2) qui se pratiquait même avec les premières technologies de l'information, pour pouvoir ensuite interagir réciproquement et participer ainsi à la mise en fonction d'un nouveau processus d'évaluation de ces systèmes : le principe de l'effet-retour et de validation de l'information reçue adopté par les cybernéticiens (Voir modèle Bertalanffy Fig.3).

1.2.5 COMMUNICATION ET INTERFACES : UN BESOIN D'UNIVERSALITÉ

C'est en réalité sur ce point particulier de rencontre qui constitue l'essence des réseaux d'information et de communication, là où se joignent l'infrastructure de communication, les services d'information et les produits documentaires, que nous orientons notre intérêt actuel. Car c'est au niveau des interfaces que les systèmes d'information intégrés se heurtèrent à une pierre d'achoppement constituée par la fragilité de ces liens qui devraient unir des modules parfaitement opérationnels séparément. Comme le signale Serge Cacaly « depuis la nuit des temps, entre l'information et l'utilisateur de nombreux intermédiaires ont toujours opéré, variant selon les époques, toujours plus ou moins réducteurs, plus ou moins intelligents »^[51]. À l'époque des nouvelles technologies de l'information, des architectures client-serveur et des systèmes distribués, la notion d'interface, dans sa signification fonctionnelle et ergonomique^[52], est reprise dans une optique encore plus profonde qui touche les différents niveaux d'interactivité entre trois composantes essentielles : « un système de communication intégré au sein du système d'information, un processus informationnel impliquant tant la transduction effectuée par le système de communication que le traitement intelligent - local ou distant - des données effectuées par l'utilisateur final »^[53] et un utilisateur final capable, sans intermédiation professionnelle, de devenir agent actif dans la conception et le développement du système informationnel dans lequel il agit.

Nous nous intéresserons particulièrement à la catégorie des interfaces Homme-Machine dans leurs formes les plus variées. La terminologie abonde de concepts du genre « Interface de communication », « Interface graphique », « Interface Windows », « Interfaces évoluées », « Interfaces en langage naturel », « Interfaces vidéotex », « Interfaces des stations de travail »..

[51] Serge CACALY. Le paysage documentaire de demain ou le triomphe des interfaces. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992 ; pp.177-182.

[52] L'ergonomie est en réalité un critère fondamental d'un usage réussi de tout système. Ceci sous entend la recherche d'une meilleure harmonie entre une fonction, un matériel et son utilisateur.

[53] Roland DUCASSE. Modèle Client-Serveur et systèmes d'information distribués : éléments pour une réingénierie des services d'information scientifique et technique / Roland DUCASSE. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. p.136.

qui s'érigent en tant qu'intermédiaires « intelligents » de codification et décodification transparentes des données dans un mouvement transactionnel entre utilisateur et système.

Par Interface Homme-Machine (IHM), il est communément fait référence à des états simples ou complexes d'interaction entre l'homme et la machine dans un processus de manipulation de données (saisie, traitement, diffusion). Dillon définit l'interface comme « masque » offert par la machine à l'utilisateur pour pouvoir accomplir les tâches suivantes de manière réversibles : (a) l'utilisateur peut communiquer ses requêtes et réponses à la machine dans un processus multimodal ; (b) la machine peut fournir les données à l'utilisateur dans plusieurs formes de présentation conviviales et intuitives^[54]. Derrière cette définition simpliste des IHM, qui retransmet le côté uniquement visuel et frontal de l'interface, résident en fait des aspects de fonctionnement complexes dues à l'interférence d'autres programmes et protocoles des domaines informatiques et des télécommunications qui constituent, selon Roland Ducasse, le domaine réel où l'intervention professionnelle se justifie face à la désintermediation progressive des phases de communication et d'usage d'information. « *L'intervention professionnelle, on l'aura compris, est, à nos yeux, surtout nécessaire en amont, dans l'ingénierie de conception et de développement des interfaces usagers-systèmes, des interfaces de communication entre systèmes hétérogènes, dans la définition de normes et de standards, le balisage, le calibrage ou dimensionnement des vecteurs, le conditionnement, le marquage, la gestion des supports* »^[55]. Car, et dans la même optique, la fonction clé d'un système d'information distribué, est la fonction de communication qui s'établit à travers l'interface intelligente en position de relais entre une source et un destinataire.

Or, l'utilisation d'outils de plus en plus nombreux et diversifiés dans leurs fonctionnalités a créé un besoin urgent pour une homogénéisation de l'interfaçage. Comme l'expose L.ABBA Et. al. , « *les problèmes essentiels qui se posent dans le projet et l'élaboration d'un Réseau Informatique.. sont.. fortement liés à la nécessité de trouver une solution interopératoire et qui fonctionne avec les diverses composantes présentes dans les lieux auxquels se raccorder. Chaque lieu a une histoire qui lui est propre, en rapport avec des choix antérieurs sur le plan technique et avec les stratégies passées et présentes en matière de financement* »^[56]. Ce problème a fait surgir la notion de normalisation à l'avant de la scène internationale dans une tentative de définir des langages et des protocoles communs de communication. C'étaient les recommandations CCITT du genre X25 et 3X, la norme ISO/OSI et ses fameuses sept couches, puis le tout puissant TCP/IP de l'Internet (héritaire de l'ancien TCP/IP d'ARPANET), qui s'établissent progressivement comme *lingua-franca* des réseaux mondiaux de transmission de données. L'hétérogénéité des systèmes enfin admise dans les systèmes distribués, ne constitue plus un handicap devant la connectivité. Au contraire, elle devient un facteur de richesse permettant à différents serveurs d'échanger, d'interagir et de partager dynamiquement toutes sortes de programmes, données et applications hétérogènes.

En réalité, la dimension réelle de la problématique de l'interopérabilité de systèmes et l'uniformité d'interfaçage, prend ses dimensions réelles dans des contextes qui accusent une fragilité des systèmes télématiques comme c'est par exemple le cas dans des pays en voie de développement. Nous faisons référence ici au cas de produits d'information à caractère

[54] Martin DILLON. Interfaces for information retrieval and online systems : the state of the art. New York, London : Greenwood Press, 1991. 337 p.

[55] Roland DUCASSE. Modèle Client-Serveur et systèmes d'information distribués : éléments pour une réingénierie des services d'information scientifique et technique. Op.cit. ; p.137.

[56] L. ABBA etc. Principe pour l'établissement d'une interconnexion entre les réseaux informatiques des pays africains. Actes de la réunion de lancement du réseau d'échanges d'informations entre institutions en Afrique. Dakar, Sénégal. 27-29 février 1992.

professionnel ou scientifique qui fonctionnent en mode monoposte ou en réseau local (parfois en mode bilingue associatif : cas des systèmes d'information bilingues arabe/latin) sans pour autant parvenir à s'établir en une sorte de produit informationnel (i.e. des OPACS^[57]) pour une communauté universelle, décentralisée et variée. Une couche interface standardisée fait incontestablement défaut à un grand nombre de systèmes, localement opérationnels et satisfaisants mais qui dans une dimension internationale deviennent inaccessibles donc forcément inconnus et non recensés.

Ce que nous observons aujourd'hui dans un grand nombre de pays en voie de développement c'est que chaque secteur cogite indépendamment dans son coin sans se soucier des problèmes des autres. Les télécommunications se débattent dans les problèmes sans fin de la téléphonie vocale ; les informaticiens s'isolent dans leurs laboratoires face à des problèmes de programmation classiques, les informatistes se perdent entre leurs documents et bibliothèques. Ce qui manque généralement c'est le fil conducteur qui permet de créer la complémentarité des tâches entre un informatiste qui collecte et gère les données, un informaticien qui structure les supports électroniques des données et conçoit les interfaces nécessaires à leur diffusion en local et à distance, et enfin un ingénieur en télécommunications qui assure leur transmission sur les réseaux nationaux et internationaux.

À ce niveau, le rôle des normes se fait lourdement sentir dans un champ d'activité qui constitue une alternative capable d'impliquer des partenaires du monde entier dans un processus d'accès, d'échange et de partage de l'information : les OPACs et la recherche en ligne. C'est en quelque sorte la technique la plus ancienne et la plus apte à assurer le concours des pays en voie de développement dans le contexte universel des phénomènes informationnels du genre : « Flux Transfrontières de Données », « Accès Universel aux Publications », « Contrôle Bibliographique Universel » et bien d'autres qui rappellent toujours les efforts déployés pour la mise en œuvre de ce mécanisme unificateur des méthodes et procédures informationnelles à l'échelle mondiale. Ceci nécessite a priori, une reconsidération de certaines normes internationales d'interfaçage. La normalisation du traitement des données bibliographiques (CCF, MARC...), les réseaux d'échange des données (EDI, UN/EDIFAC...), les navigateurs et les moteurs d'interrogation des bases de données (SGML, NETSCAPE, WAIS...) doivent confirmer leur mondialisation non pas par l'imposition de modèles standards mais en prenant en compte les particularités de la communauté internationale, en essayant de répondre aux exigences d'ordre structurel, linguistique, culturel, économique ... de tous les partenaires concernés.

Ceci dit, « la caractéristique qui dépasse le simple mariage de l'information et de la communication est celle qui permet à des systèmes hétérogènes de communiquer entre eux, à effectuer des échanges grâce à des interfaces conviviales »^[58] et surtout dynamique. Car, il ne s'agit plus de créer l'artefact qui s'établit en intermédiaire passif entre une source et une destination. Face à une extension considérable des espaces géographiques de communication et ses débouchés sur des aires culturelles variées, l'IHM est censée avoir, en plus de son rôle médiateur, un soubassement culturel intense qui adhère aux particularités cognitives de l'interprétation et de la compréhension des sens ainsi qu'aux opérations de représentation et codification qui accompagnent tout processus de communication. Ceci fait automatiquement appel à des domaines annexes du traitement des connaissances tels que l'intelligence artificielle et les systèmes experts ; les sciences cognitives et les industries de la langue dans tous leurs aspects techniques : traitement automatique du langage naturel, traduction assistée par

[57] On-line Public Access Catalogue

[58] Mohamed BEN AHMED. Création culturelle et nouvelles technologies de l'information. Al Madar (Cité des sciences. Tunisie). N°3, 1995 ; pp.21-32

ordinateur, reconnaissance optique des caractères, traitement de la parole...communément identifiées comme relevant de l'ingénierie linguistique ou du génie linguistique.

1.2.6 INTERFACES ET LANGAGE : LES INDUSTRIES DE LA LANGUE

Le concept d'industrie de la langue s'est situé, dès son apparition en 1984^[59], dans la perspective plus vaste des Nouvelles Technologies de l'Information qui puisent leurs repères dans une logique économique de productivité et de compétitivité à l'échelle mondiale.

Partant du dicton américain qui prévoit que « l'enjeu est désormais de donner la bonne information, à la bonne personne, au bon moment, pour pouvoir prendre la bonne décision et gagner la première place sur les marchés », l'innovation technologique, qui constitue un passage obligé pour le contrôle des marchés dans un contexte de concurrence, nécessite une activité de recherche et de développement consommatrice par excellence d'information spécialisée. À part le problème des réseaux de télécommunications internationaux qui imposent des standards de fait et des langues de communication propres à certains des pays développés, l'implication de partenaires économiques de profils linguistiques différents et disposant de sources locales d'information économique, scientifique et technique généralement bilingues ou multilingues, fait face à une problématique linguistique d'interfaces que seuls les produits des industries de la langue peuvent résoudre.

Dans son approche économique du rôle des industries de la langue, Rondeau confirme que *« l'examen de certains facteurs qui exercent une influence économique fait apparaître clairement l'importance du rôle des industries de la langue dans un contexte multilingue : celui de doter la communauté mondiale d'instruments technologiques nouveaux qui permettraient de traîner les diverses langues, d'ouvrir ainsi les canaux de la communication mondiale et éventuellement de s'affranchir des barrières linguistiques »*^[60].

Tous ces phénomènes attribués à l'industrie linguistique et au bilinguisme d'interfaçage, ont été largement abordés dans leurs aspects conventionnels par une grande partie de la littérature scientifique occidentale. Il s'agit toujours d'un (bi)multilinguisme que nous considérons « souple » car il se pare d'une interopérabilité entre deux ou plusieurs systèmes linguistiques qui, s'ils diffèrent sur un plan morphosyntaxique, restent uniformes, jusqu'à certaines limites de compatibilité, au niveau de l'acte scripto-graphique et directionnel qui utilise les mêmes tables de codification et les mêmes procédures de représentation et d'interprétation des données. Un (bi)multilinguisme « lourd » impliquerait, en plus des considérations morphologiques, syntaxiques, sémantiques et pragmatiques propres à des langues d'origines similaires, des difficultés supplémentaires relatives aux caractères et à leurs extensions de code (par rapport à l'ASCII standard) ainsi qu'aux sens de la rédaction (de la droite vers la gauche et du haut vers le bas) qui compliquent le principe de la représentation simultanée de textes diamétralement opposés dans le sens graphique sur des interfaces bilingues.

Ces particularités scripto-graphiques laissent entrevoir le parcours encore complexe que des sociétés en voie de développement caractérisées par un multilinguisme « lourd », vont parcourir pour atteindre des degrés de développement des interfaces bilingues enregistrés dans les milieux plus favorables du bilinguisme « souple ». Avec la jonction de cet aspect purement technique aux aléas des interférences sociopolitiques et culturelles auxquelles il est constamment soumis, des

[59] Jean-Claude RONDEAU. Multilinguisme et génie linguistique. Université, mai-juin 1993 ; p.46.

[60] Jean-Claude RONDEAU. Multilinguisme et génie linguistique. Op.cit. ; p.47.

grandes questions d'opportunité scientifique, d'acculturation de la science et de la recherche, de l'identité culturelle et de l'autodétermination, font sentir leurs empreintes conflictuelles dans les discours politiques et les programmes de développement socio-économiques.

Dans ces sociétés en voie de développement, deux approches contradictoires font valoir leurs propres hypothèses face à cette problématique technolinguistique.

D'une part, la communauté scientifique et de recherche prêche pour l'acculturation du savoir humain et de la science afin de justifier une appartenance inconditionnée à la dynamique de la société d'information universelle malgré le statut de partenaire à sens unique (consommateur) qui leur y est attribué. La devise des scientifiques et chercheurs prévoit que les problèmes propres au sous-développement, les caractères particuliers de la technologie et la structure du marché font que les solutions satisfaisantes doivent être recherchées à l'échelle de la coopération internationale. D'ailleurs l'histoire des inventions a toujours montré que l'évolution des sciences et des techniques s'est accomplie grâce aux refontes que certains ont effectuées sur les vestiges des autres sans critères de limitation géographique ni de zèle culturel ni de chauvinisme idéologique ni de durcissement politique. La langue devient dans ces cas un facteur aculturel qui obéit aux seuls critères de l'intérêt scientifique, économique et technique.

Face à ce langage universaliste, un langage plus coercitif, celui de la logique subjective des décideurs politiques s'engage dans une nécessité de réadaptation des technologies en vue de préserver ce qu'on appelle « *l'identité culturelle* ». Alimenté par un échec des projets NOMIC et par une mondialisation rampante des modèles socio-économiques dominants, le langage politique laisse entendre un retour vers la pensée nationaliste et le recours aux regroupements régionaux porteurs de valeurs unifiées.

Seulement, si cette logique trouve ses arguments dans une dimension socioculturelle généralement admise même dans les sphères de compétition entre pays développés, la question fondamentale ne concerne pas le recours à une langue ou à une autre pour développer des secteurs d'activités stratégiques. Il est plutôt question de savoir, en face d'une universalisation du savoir humain, d'un manque des capacités techniques adaptées à des critères linguistiques particuliers et en l'absence d'un patrimoine de littérature scientifique et technique locale et compétitive, s'il est plus logique de continuer le processus d'interactivité avec le contexte international dans des langues imposées ou faudrait-il attendre de régler les plates-formes linguistique et technologique locales pour s'insérer dans le cours de la dynamique internationale des échanges d'IST ?

Quoi qu'il en soit, un choix ou un autre ferait certainement face à la contrainte majeure du facteur temps. Il nous semble totalement irrationnel, au moment de la course contre la montre pour l'exclusivité informationnelle et la primauté dans les inventions, de prévoir une politique de redynamisation technologique basée sur la notion du recul « stratégique » ; celle qui consiste à laisser l'initiative d'expérimentation aux autres et de prendre les solutions clés en main comme raccourcis stratégiques de récupération. C'est là en fait un phénomène qui s'est établi, entre autres, dans le contexte du (bi)multilinguisme des IHM. Dans un contexte comme celui de la langue arabe, la plate-forme Windows arabe, les navigateurs Web de l'Internet en multilingue (y compris l'arabe) ... sont des exemples concrets de cette léthargie de l'initiative locale des cultures concernées pour créer l'alternative linguistique qui les concerne. Ce sont généralement des concepteurs de logiciels étrangers à la langue en question qui, souvent guidés par une politique de marché, créent l'alternative manquante. Or, comme le précise Rondeau « *il faut connaître la langue pour pouvoir la traiter avec l'aide des machines. Cela suppose une meilleure connaissance des phénomènes linguistiques et une augmentation du savoir technique*

pour que les produits et les services soient de plus en plus adaptés en fonction des besoins des utilisateurs »^[61]. S'il est vrai que ces produits ont été conçus grâce au concours de spécialistes d'origine linguistique arabe, il est aussi vrai que ces spécialistes sont généralement formés hors du « *carcan de l'orientalisme traditionnel* », donc libres de prendre des initiatives qui échappent souvent à la réalité socioculturelle vécue et par conséquent conduits vers le développement de nouvelles conceptions ou, plus modestement, et plus souvent, vers l'adaptation à l'arabe des produits développés par l'analyse formelle d'autres langues^[62].

Les contre-exemples à ce point de vue plaident en faveur d'une dissociation épistémologique entre les notions contenant-media-contenu pour parvenir à une internationalisation des apports des sciences et des techniques au profit de l'humanité entière. Dans ce sens, les exemples ne font pas défaut. Les ressources d'IST internationales sont classées dans un ordre qui donne aux langues anglaise et française une priorité linguistique évidente (rapports volume/langue). Ceci n'a pas empêché les pays à domination linguistique moyenne et faible (langues latines à part le français et langues asiatiques) de recourir aux ressources anglophones et francophones pour y puiser les bases fondamentales de leur développement socio-économique. Un autre contre-exemple se situe au sein de la Communauté Économique Européenne, là où une tendance naturelle a pris forme à n'utiliser qu'une ou deux langues parmi les plus répandues pour faciliter les pratiques commerciales.

Où se situe donc la réalité du vrai problème concernant le multilinguisme des interfaces en arabe/latin pour l'accès à l'IST? Est-ce un problème technologique uniquement ou est-ce un phénomène de prédisposition culturelle, économique et politique pour maintenir le rythme de l'évolution internationale en matière des sciences et des techniques ?

2. PROBLÉMATIQUE

La réponse à tant de questions ne saurait à notre sens échapper à une segmentation supplémentaire du domaine du multilinguisme pour tracer des limites de spécialité endogènes au secteur du traitement multilingue de l'information entre deux domaines particuliers : l'analyse de contenu et le traitement du rendu ou de la représentation.

Le domaine du traitement linguistique peut couvrir en effet deux sous-domaines de recherche : d'une part le traitement syntaxique et sémantique du contenu, i.e. la signification des valeurs phonétiques de l'information, les modalités d'accès aux documents structurés... et d'autre part la représentation graphique et visuelle du contenu. Chaque domaine placera des valeurs différentes dans le traitement de l'information textuelle concernée. La recherche d'information textuelle dans une base de données nécessiterait des techniques complètement différentes de celles requises pour la représentation graphique des mêmes données. Dans le premier cas le traitement concerne plutôt le contenu des données sans s'occuper de la forme de leur présentation. La définition de la structure d'un document en format SGML, ODA, Hytime... est un exemple d'une procédure de traitement de contenu qui reste indépendante des particularités relatives à la police de caractères. Par contre, le deuxième cas serait plutôt enclin à aborder les effets d'apparence et de présentation visuelle au cours de laquelle les ressources graphiques (fontes, polices, techniques de codage...)

[61] Jean-Claude RONDEAU. Multilinguisme et génie linguistique. Op.cit. ; p.46.

[62] D.E.KHOULOUGHLI. Linguistique arabe et informatique : recherches actuelles dans les universités de Paris. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. p.4.

seraient très demandées. C'est en quelque sorte le domaine des interfaces graphiques et des méthodes d'affichage synoptique^[63].

De la détermination de ces deux approches dans le traitement des données textuelles, deux domaines de recherche sont actuellement développés en bifurcation d'une discipline générique, en l'occurrence le traitement linguistique de l'information. Le premier est tributaire du domaine de l'analyse structurelle des documents et des méthodes de repérage : interfaces de recherche, interrogation en langage naturel, structure de contenu... Le deuxième domaine dont l'axe central est le format de la représentation du contenu (graphic rendering) couvre l'aspect ergonomique et interactif avec l'utilisateur final à travers la convivialité de la couche interface supérieure des applications : format d'écran, organisation spatiale des données, densité surfacique et informative, multifenêtrage, gestion des couleurs et des contrastes, infographie et codage des données ...

Dans cet enchevêtrement de concepts techniques, notre intérêt portera, comme nous l'avons déjà annoncé, sur l'un des éléments qui constituent la jonction entre les trois protagonistes rassemblés (information, technologie, communication), en l'occurrence les Interfaces automatisées de dialogue Homme-Machine dans un contexte aussi instable et aléatoire que celui des pays de langue arabe. C'est en réalité le point névralgique au niveau duquel se définissent les opérateurs réels de la réussite ou l'échec de l'action d'intégration des forces de la recherche et de la communication scientifique.

Nous nous limiterons dans cette étude à l'analyse de la représentation graphique plutôt que de la structuration linguistique du contenu. Nous aborderons cette approche du point de vue de l'interopérabilité des systèmes au niveau des différentes phases de traitement technologique, de l'information et de la communication et des solutions fournies et utilisées de part et d'autre (producteurs de ressources, pourvoyeurs de services d'accès, concepteurs de médias, utilisateurs de données) pour répondre aux exigences informatives et communicationnelles d'un contexte bien particulier, celui des environnements multilingues arabo-latins. Car nous croyons, comme l'observe DALBIN, que « *l'évolution conjuguée des utilisateurs et de leurs besoins d'une part (la demande), du flux et des volumes d'information accessibles à travers des matériels informatiques et de télécommunication toujours performants (l'offre) explique le développement, ces dernières années, du concept d'interfaces intelligentes, par la nécessité d'améliorer la qualité de l'offre globale contenant, contenu et accès au contenu* »^[64].

2.1 PREMIER AXE DE RECHERCHE

La première interrogation posée sera relative aux solutions techniques apportées à des strates différentes du processus du traitement de la langue arabe pour surpasser l'obstacle purement scripto-graphique. La langue arabe, comme support d'information scientifique et technique est appelée à marquer sa présence dans l'univers informationnel et communicationnel mondial qui s'installe depuis les simples pratiques de traitement informatique en monoposte jusqu'aux fonctionnalités plus complexes d'Internet (messagerie électronique, WWW, FTP, Telnet...), de la gestion électronique des documents, de la reconnaissance optique des caractères... comme instrument véhiculaire d'une information écrite, elle est censée rester, comme toute langue de

[63] Christophe KOLSKI. Ingénierie des interfaces Homme-Machine : conception et évaluation. Paris : Hermès, 1993.

[64] Sylvie DALBIN. Interfaces dans les systèmes d'aujourd'hui. In : Christian Bornes (dir.), Les Interfaces Intelligentes dans l'Information Scientifique et Technique. Cours INRIA, 1992. p.142

communication, au diapason des évolutions que connaissent les systèmes d'information et de communication en vigueur.

En effet, face à cette expansion et dominance des systèmes ouverts et distribués, comment la partie lourde du multilinguisme (formes typographiques et direction de l'écriture), par analogie à la partie souple (syntaxe et sémantique), trouve-t-elle progressivement sa voie dans les systèmes hybrides actuels d'information et de communication ? S'agit-il d'un fait de dépendance verticale d'une langue source dominante sur une langue cible ? Est-ce plutôt une correspondance horizontale entre deux ou plusieurs langues qui s'alternent ou « cohabitent » au niveau de la même interface ? Dans quelles conditions les solutions techniques déjà apportées au multilinguisme d'interfaçage opèrent-elles ? Associatives (présence simultanée au niveau de l'interface) ou dissociatives (par alternance) ? Ceci nous conduirait automatiquement à aborder cet axe par une introduction au contexte général dans lequel les systèmes d'information et de communication puisent leurs ressources élémentaires au niveau de trois vecteurs de recherche que nous avons jugés importants : les théories linguistiques et la sémiologie, le traitement du signal et codage binaire, et l'algorithmique ; trois facteurs que nous avons jugés pertinents pour agir sur le volet fondamental et important de l'adaptation de la langue arabe au domaine du traitement informatique (codage de l'information et de sa manipulation physique : saisie, interprétation, transformation, stockage, restitution, échange, reproduction...).

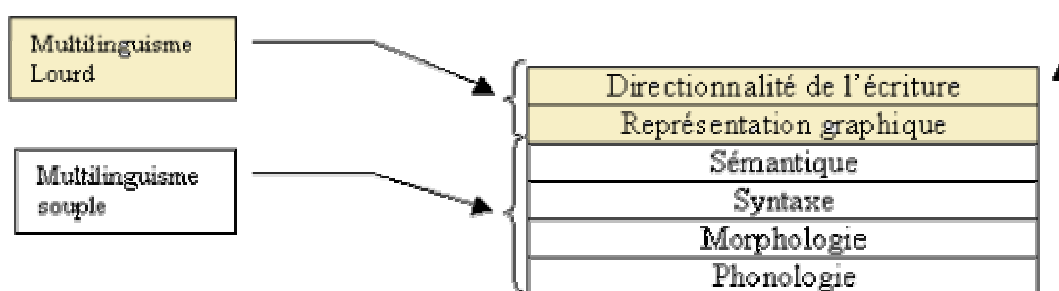


Tableau 2 : Strates de la structure linguistique d'une entité multilingue

2.2 DEUXIEME AXE DE RECHERCHE

La deuxième interrogation que nous aborderons est relative à l'aspect d'utilisabilité de la langue arabe au niveau des interfaces multilingues en arabe/latin. L'objectif de ce questionnement est de soulever les problématiques des différentes formes d'inadaptation entre cette catégorie d'interfaces et les utilisateurs potentiels.

Nous avons été sensibles tout au long de nos travaux de recherche à l'absence d'une littérature abondante et spécialisée axée sur la culture informatique arabe ; sur le questionnement des solutions informatiques proposées actuellement comme consensus implicite et leur confrontation avec les standards de fait de la langue, de ses modes d'usage tels qu'ils sont culturellement établis dans la perception collective et cognitive de ses usagers potentiels.

Généralement conçues par des concepteurs étrangers à la langue ou du moins partiellement connaisseurs des règles linguistiques, calligraphiques et culturelles de la langue arabe, les IHM arabe/latin présentent à notre sens un degré assez important d'incohérence et d'inadéquation entre le système multilingue et l'utilisateur arabophone. Ces incohérences peuvent avoir à notre sens de grandes répercussions sur l'acceptabilité, l'utilisabilité et l'efficacité même de l'interface. Par la soumission de ces solutions ad hoc aux mécanismes linguistiques de fait et d'usage, nous soulèverons beaucoup d'anomalies dont certaines sont justifiées par le référent socioculturel des

usagers et d'autres sont induites par l'influence et l'interférence de disciplines connexes (sociolinguistique, transfert de technologies, politiques culturelles...).

3. OBJECTIFS DE L'ÉTUDE

Répondre aux interrogations précédentes n'indique pas pour autant que l'on s'éloigne des standards de fait en matières d'interfaces d'accès et de diffusion de l'IST atteints par les recherches des écoles qui ont évolué dans ce domaine et qui ont tracé les étapes par lesquelles on passe pour concevoir et évaluer une interface Homme-Machine. Car il y a toute une série de standards, de règles à respecter pour construire une interface utilisateur et tester son degré d'ergonomie et d'acceptabilité par les usagers. Il s'agit-là entre autres d'une appréciation des deux critères d'utilité et d'utilisabilité fixés par GRISLIN et KOLSKI comme « *propriétés pertinentes pour le diagnostic de la qualité ergonomique* » de l'IHM^[65].

Notre objectif s'oriente sans doute vers cette optique d'évaluation non pas pour reprendre l'état de l'art en la matière, mais plutôt pour appliquer cet état de l'art sur des interfaces (bi)multilingues qui dominent le contexte d'usage de l'information scientifique et technique. Notre champ d'action restera par conséquent rattaché aux solutions logicielles d'une grande envergure d'usage car ; à notre sens, c'est là que se profilent les changements et les évolutions profondes des interfaces Homme-Machine. C'est également à travers cette gamme d'interfaces Homme-Machine que se confirment les valeurs socioculturelles et économiques des actes d'information et de communication.

Seulement, puisque la technologie, comme nous l'avons déjà signalée, est un vecteur de suprématie d'une culture d'origine, en l'occurrence celle du monde à dominante linguistique latine, l'information ne saurait échapper à la contrainte qu'exerce la langue du support sur la langue du contenu. Ceci explique les retards observés dans les domaines de la codification électronique des données, de la normalisation des procédés de transcription et d'échange et de l'industrie de la langue orientées vers les langues d'origines non latines.

Notre objectif, dans ce cas particulier, est de démontrer à quel point cette « fatalité » de dépendance à été surmontée (nous dirions plutôt contournée), pour permettre à des langues vernaculaires comme l'arabe, apparemment inadaptées (sur un plan technologique), d'accéder progressivement à l'univers électronique des systèmes d'information et de communication distribués.

Cependant, si le dit contexte relève en apparence d'un paradigme technoculturel évident (Nouvelles Technologies d'une part et Information de l'autre), une large partie du handicap que connaît la langue arabe au niveau de son usage électronique incombe à la réalité des usages qu'on en fait. Comme nous le montrerons au courant de ce travail, des considérations d'ordre culturel et historique agissent de manière très évidente sur la nature et les limites du bilinguisme actuel.

[65] Martial GRISLIN, Christophe KOLSKI. Évaluation des interfaces Homme-Machine lors du développement des systèmes interactifs. Technique et Science informatiques. Vol.15-n°3/199. p.266

4. MÉTHODOLOGIE DE TRAVAIL

Notre propos, portant sur la thématique des interfaces bilingues arabe/latin dans l'accès et la communication de l'IST, sera présenté selon ses quatre étapes principales :

- 1) Dans une première partie : un préliminaire sur le contexte général de l'information et de la communication pour mettre en exergue les facteurs qui agissent directement sur ses divers aspects fonctionnels : les théories linguistiques et la sémiologie dans leurs rapports intrinsèques avec les théories de l'information et de la communication ; la théorie du signal comme vecteur important dans le codage et la transmission de l'information électronique ; les techniques de codage et de décodage dans les processus de la représentation et de l'analyse de l'information ; l'algorithmique comme processus intelligent dans la négociation et la désignation des opérandes du travail multilingue (analyse de contexte, tri multilingue, bidirectionnalité...). L'objectif visé par la mise en relief de ces facteurs est de procéder par une approche progressive descendente pour cerner le cadre de développement de la problématique centrale : jauger la profondeur des réalisations qui ont marqué les efforts de l'intégration de la langue arabe en informatique à travers la multilingualité des IHM dans un contexte informationnel arabe/latin.
- 2) Dans une deuxième partie, nous aborderons le principe du codage des données comme continuation logique vers le traitement informatisé de la langue arabe. L'une des confusions majeures actuelles pour les profanes de l'informatique est due à la multiplicité linguistique des ordinateurs et des progiciels d'usage dans les domaines Bureautique et spécialisé. Le tout-anglais, ou du moins le tout-latin est une réalité indiscutable face à « l'intrusion » parfois contestée des langues non latines dans l'interaction avec la machine. L'objectif de cette partie est de procéder à une approche comparative des solutions de codage de l'information sur des principes linguistiques universels. Le problème de forme des caractères n'est qu'une partie insignifiante de l'enjeu linguistique. Tout passe par la désignation des tables des codes de caractères que les systèmes mettent en action pour « forcer » une machine à afficher un caractère ou un autre. En termes d'échange, il est plutôt question d'accord entre émetteur et récepteur quant à l'harmonisation des plates-formes de travail identiques plutôt que d'attribuer le handicap à l'inadéquation d'une langue ou une autre au principe du partage des données. Une grande partie de l'analyse dans cette partie sera réservée au monde actuel d'Internet et des systèmes ouverts et distribués. L'internationalisation est aujourd'hui un axe de recherche qui trouve en Internet le champ d'action par excellence pour tester, installer et évaluer sa créativité linguistique au niveau des systèmes ouverts et distribués et des interfaces Homme-Machine.
- 3) La troisième partie de ce travail progressera en profondeur dans les systèmes d'information multilingues arabe/latin pour étudier les aspects d'usage de la langue arabe du point de vue général de traitement et d'échange de l'information scientifique et technique. Depuis les plates-formes monopostes sous DOS et Windows jusqu'aux toutes nouvelles applications en réseau sous Internet, la langue arabe a toujours nécessité un traitement particulier au niveau de la représentation graphique de ses caractères ainsi que de la directionnalité droite-gauche de son système d'écriture (la partie lourde du multilinguisme). Les documents composites et les systèmes partagés des réseaux distribués constituent une difficulté supplémentaire que les systèmes monopostes n'avaient pas affrontée massivement. La négociation et l'étiquetage de la langue dans un univers hypertextuel et virtuel comme celui d'Internet constitue le véritable défi actuel pour les langues vernaculaires y compris l'arabe. Quel état ? Quel bilan ?

- 4) La dernière partie visera à exposer les problématiques d'usage de toutes les solutions proposées. La juxtaposition entre l'état de l'art en termes de solutions bilingues concernant l'arabe sera confrontée à l'état de l'art d'une langue qui paraît parfois se contenter d'accepter la solution informatique développée dans des sphères culturelles qui lui sont étrangères. Ces anomalies seront traduites par des exemples puisés au niveau de certaines interfaces Homme-Machine arabe/latin d'usage actuel. L'objectif de ce questionnement est avant tout épistémologique : soulever une problématique de recherche rarement exposée au moment de la conception de produits multilingues pour lesquels la langue arabe est prise comme partenaire linguistique.

Ces exemples seront traités en fonction de deux aspects fondamentaux : d'une part leur structure fonctionnelle (similitudes et diversités) et d'autre part leur adaptabilité aux exigences pratiques de la recherche et leur acceptabilité auprès des utilisateurs dans divers contextes d'usage.

Ceci implique automatiquement le recours à des méthodes rigoureuses d'évaluation d'IHM. Nielsen et Molich avancent la théorie des approches dites prédictives et des approches dites expérimentales^[66] alors que Senach établit une distinction entre approches analytiques et approches empiriques^[67] ; Sweeney et Maguire proposent quant à eux trois méthodes basées sur l'utilisateur, la théorie et l'expertise^[68] ; Whitefield et ses collègues organisent leur classification en fonction de la présence ou non de l'utilisateur d'une part, et du système interactif à évaluer d'autre part^[69] ; Coutaz et Balbo^[70] centrent leurs techniques sur plusieurs paramètres tels que le degré d'automatisation, la prise en compte de l'utilisateur, ou encore le savoir utilisé pour l'évaluation.

Notre approche sera quant à elle inspirée essentiellement de la classification élaborée par GRISTIN et KOLSKI qui repose sur trois catégories d'approches fondamentales :

- ❑ Les approches centrées sur les utilisateurs ;
- ❑ Les approches centrées sur une expertise humaine ou « papier » ;
- ❑ Les approches centrées sur une modélisation de l'IHM et/ou de l'interaction Homme-Machine.

5. PRECISIONS TERMINOLOGIQUES

Il est très courant de se retrouver face à un concept que l'on s'imagine maîtriser parfaitement, voire sur lesquels on ne se pose nullement de question quant à leur portée sémantique. La langue est truffée de pléonasmes et de faux amis. Les emprunts linguistiques, les consensus d'usage, les translittérations... constituent autant des sources de flottement dans la définition de certains concepts. À titre d'exemple, des termes dans le champ d'action de notre étude apparaissent dans les usages courants comme des synonymes : bilinguisme et multilinguisme ; interface et interface

[66] J. NIELSEN, R. MOLICH, heuristic evaluation of users interfaces, Actes de colloque CHI'90, Seattle, ACM New York, pp.349-356, 1990

[67] B. SENACH, P. ALENGRY. Évaluation d'un dispositif d'assistance à la conduite et au dépannage de chaînes de fabrication automatisées. rapport INRIA n°163. Mars 1994.

[68] M. SWEENEY et al. Evaluating user-computer interaction : a framework. International Journal Man-Machine Studies, 38, 689-711, 1993.

[69] WHITEFIELD et al. A framework for human factors evaluation. Behavior & Information Technology, vol. 10, n°1, pp.56-79, 1991.

[70] J. COUTAZ & S. BALBOS. Évaluation des interfaces utilisateur : taxonomie et recommandations. Actes des sixièmes journées sur l'ingénierie des interfaces Homme-Machine, IHM'94, 211-218, Lille, 8-9 décembre 1994.

intelligente ; combinaison binaire et code binaire ; Unicode et ISO/IEC 10646 ; jeux de caractères codés et tables des codes, etc.

Pour garantir une homogénéité terminologique, nous avons jugé important de préciser la signification exacte que nous donnons à certains concepts et termes d'usage dans le domaine du bilinguisme (et multilinguisme). Notre choix a porté sur les termes et concepts suivants qui constituent le soubassement de notre travail actuel.

Les jeux des caractères codés

Un jeu de caractères est par défaut un ensemble de caractères définis pour le traitement d'une ou de plusieurs langues. Bien qu'un jeu de caractères soit généralement conçu pour une seule langue, plusieurs peuvent partager le même répertoire de caractères codés d'une langue quelconque. Le jeu de caractères codés selon la norme ISO-8859-1 peut à titre d'exemple être utilisé pour la majorité des langues de l'Europe de l'Ouest. Le jeu des caractères arabes normalisés ASMO 449 est utilisé pour le traitement de plus de dix langues (arabe, persan, afghan, swahili, ourdou, berbère ...).

Le jeu de caractères est indépendant de son propre encodage. L'encodage d'un jeu de caractères est une méthode qui permet de convertir un jeu de caractères dans une forme lisible par ordinateur pour des motifs de transfert ou de stockage de données. Souvent un jeu de caractères est codé sous plusieurs formes. À titre d'exemple, l'un des jeux de caractères utilisé pour représenter la langue japonaise est le JIS X208 qui peut être représenté en trois modes différents : le Shift-JIS, le EUC-JP et le JIS. Cette diversité est souvent tributaire des plateformes de travail utilisées.

Bilinguisme et multilinguisme

Il est évident que de par la structure morphologique des deux concepts, la nuance qui les distingue est implicite. Le « Bi » sous-entend « deux » et le « Multi » suggère le nombre « deux » ou plus. Or l'usage est tel que les deux termes s'utilisent confusément. Cette source de confusion est essentiellement issue des influences d'ordre culturel, géographique et linguistique.

Selon le dictionnaire Larousse, le bilinguisme est la « *pratique de deux langues (par un individu ou une collectivité)* »^[71]. Les deux langues peuvent être de la même famille (français - espagnol) ou de familles différentes (français - arabe). Les deux exemples du bilinguisme canadien (français et anglais) et le bilinguisme maghrébin (arabe et français) expliquent cette particularité linguistique.

Le multilinguisme en revanche correspond à la « *situation d'une région, d'un État, etc. où sont parlées plusieurs langues* »^[72]. Le multilinguisme est synonyme de plurilinguisme.

La confusion au niveau de l'usage d'un concept est généralement observée au moment de la confrontation de deux familles linguistiques. Entre la langue arabe et les langues latines, les nuances sont faibles car elles dépendent du côté où l'on se situe du contexte dans lequel on évoque le concept et le public auquel on s'adresse. En effet, un européen parlera plutôt de bilinguisme que de multilinguisme car il juxtapose sa propre langue (français, anglais, espagnol ou autre) avec une seule langue qui est l'arabe. Un arabe parlera plutôt de multilinguisme car la

[71] Bibliorom Larousse. Version 1.0. Microsoft Corporation & Liris Interactive. 1996

[72] Bibliorom Larousse. Op.cit.

juxtaposition s'effectuera entre une langue propre qui est l'arabe et une conglomération de langues latines des pays d'Europe avec lesquelles il traite. Le contexte d'usage influe également sur les termes à utiliser de sorte que les contextes cultes (académiques et savants) parlent généralement en termes de multilinguisme prêchant ainsi l'universalité et la transparence linguistique. La répartition géographique détermine parfois le choix d'un terme par rapport à un autre. Dans la région du Maghreb arabe, on parle de bilinguisme arabe – français ; au Moyen Orient il est plutôt question de bilinguisme arabe – anglais ; dans certaines autres zones il est plutôt question de multilinguisme arabe – français – kabyle (Algérie), arabe – anglais – copte (Egypte) ou arabe – anglais – hébreu (Jérusalem) ...

Malgré les différences qui existent entre les langues latines et germaniques (anglaise et allemande), notre étude traitera de toutes ces langues comme d'un vis-à-vis linguistique homogène à la langue arabe, car notre souci étant de traiter les difficultés d'un contexte de multilinguisme lourd ou la représentation graphique et la directionnalité de l'écriture jouent un rôle très important dans la cohabitation de deux familles (ou plus) de langues, et ce, dans une même application informatique ou dans un même produit d'information électronique.

Ergonomie

Terme composé des mots grecs *ergon* (travail, œuvre) et *nomos* (loi) désignant la science des conditions de travail et leur meilleure adaptation possible en faveur de l'homme. Les lignes directrices de l'ergonomie préconisent de concevoir les machines et les outils en respectant les conditions anatomiques, physiques et physiologiques de l'utilisateur.

Les postes de travail avec écran sont particulièrement visés, car ils induisent une position du corps défavorable à l'organisme. Un lieu de travail ergonomique suppose des propositions sur la technique, la conception, la hauteur et l'orientation du bureau, du siège, de l'éclairage et de l'écran. Les directives de l'Union Européenne, en vigueur depuis le 1/1/1997 sur l'aménagement ergonomique de tels postes de travail, prévoient également une certaine régulation des horaires. Un employé peut maintenant exiger un poste de travail conforme aux directives européennes.

Certains périphériques d'entrée connaissent en la matière un grand développement (i.e. clavier, souris). Le clavier ergonomique, comme le Natural Keyboard de Microsoft, est légèrement arrondi, ce qui correspond à la position anatomique des mains lorsque l'on écrit. Cette position naturelle est censée prévenir les maladies connues sous le nom générique de RSI (Repetitive Strain Injury), provoquées par une surcharge des nerfs, des muscles et tendons de la main. L'ergonomie du clavier Natural Keyboard de Microsoft s'adapte à la position naturelle des mains.

Outre le matériel, les logiciels sont eux aussi considérés de plus en plus d'un point de vue ergonomique (convivialité et facilité d'emploi). Les critères sont ici la possibilité d'une utilisation intuitive, l'emploi de couleurs, l'automatisation de certaines tâches, la durée de l'apprentissage et la tolérance aux pannes avant d'éventuelles pertes de données.

D'un point de vue des interfaces, Sylvie DALBIN analyse leur ergonomie en fonction des deux critères suivants :

1. **L'aspect matériel** (poste de travail) : il s'agit d'évaluer « *le temps de réponse, la qualité de l'écran, les dispositifs de communication (tels le clavier, le stylet, l'écran tactile, la souris, l'interface vocale), les dimensions ...* »^[73] ;
2. **Les logiciels** (organisation des données et stratégies d'accès aux données) : la disposition de l'information sur l'écran, les polices de caractères, la typographie du texte, la densité de visualisation, la facilité d'utilisation ...

Interface Homme-Machine

Le terme « interface » dispose d'un grand nombre de soubassements conceptuels qui convergent vers la désignation d'une « *jonction permettant un transfert d'informations entre deux éléments d'un système informatique* »^[74]. Il est certes utilisé dans des expressions comme « *interface vidéotex* », « *interface en langage naturel* », « *interface graphique* », « *interface des stations de travail* », « *interface évoluée* », « *interface de communication* », « *interface Windows* »^[75] ... L'interface Homme-Machine se distingue dans ce lot de concepts par son rapport en termes de jargon informatique à la partie visible par l'utilisateur d'une application (ex. : écrans de saisie). Ce terme s'est surtout développé avec l'avènement des interfaces graphiques et l'émergence de programmes générateurs d'IHM (ex. : HIGH SCREEN de PC SOFT), qui permet de créer des interfaces au « look Windows » sous DOS : le revamping).

Interfaces intelligentes

Le développement des IHMs est allé crescendo pour devenir un domaine qui relève de l'intelligence des systèmes d'information et des systèmes experts. Les interfaces intelligentes constituent aujourd'hui le secret de la réussite commerciale des solutions informatiques proposés au grand public.

L'intelligence des interfaces « *vient de ce que le système est adapté (il ne s'auto adapte pas encore) aux utilisateurs, et est conçu en fonction de leurs besoins, de l'approche que ceux-ci peuvent avoir de ces systèmes et de leur contenu, donc par rapport à des objectifs, des fonctionnalités à assurer, des contraintes et un environnement précis* ».^[76]

Pour ce qui relève de notre thème, à savoir les interfaces multilingues, une interface intelligente se doit de savoir négocier automatiquement le choix d'une langue dans un document multilingue composite. L'interface doit analyser la structure d'un document en phase de réception pour définir ses différentes parties et reconnaître à travers les paramètres de sa forme et de son contenu, sa composition interne. Comme nous le verrons dans la dernière partie de ce travail réservée à l'étude comparative d'interfaces multilingues, certaines opèrent un choix automatique des langues actives en fonction de la composition du document, d'autres nécessitent l'intervention de l'utilisateur pour paramétrer son environnement multilingue.

[73] Sylvie DALBIN. Interfaces dans les systèmes d'aujourd'hui. In : Christian Bornes. Dir. Interfaces Intelligentes dans l'information scientifique et technique. INRIA, 1992. pp.133-173.

[74] Petit Robert : dictionnaire analogique et alphabétique de la langue française. Nouvelle édition, version électronique. 1996.

[75] Sylvie DALBIN. Interfaces dans les systèmes d'aujourd'hui. Op. Cit.

[76] Sylvie DALBIN. Interfaces dans les systèmes d'aujourd'hui. Op. Cit.

6. REVUE DE LA LITTÉRATURE

Notre choix pour établir la liste des références bibliographiques à valeur ajoutée à notre travail, est inspiré de deux objectifs essentiels : d'une part nous comptons présenter les travaux antérieurs qui ont porté sur ces questions et auxquels nous apporterons soit une confirmation soit une divergence de point de vue ; d'autre part, nous contribuerons à travers cette revue de la littérature à l'orientation vers des travaux dont la thématique reste très attachée à notre problématique générale sans pour autant entrer dans le noyau central de notre étude. C'est le cas de tous les travaux sur l'analyse linguistique de la langue arabe qui constituent un complément très proche de notre étude orientée plutôt vers l'aspect scripto-graphique arabe dans les systèmes d'information multilingues.

Nous introduirons également au terme de cette revue de la littérature une liste d'URLs qui pointent vers des institutions pilotes spécialisées dans l'internationalisation d'Internet et tout particulièrement les développeurs de solutions multilingues en arabe/latin. Nous considérons ces sites comme une source très importante d'information qui permet d'accomplir une veille stratégique sur l'évolution de ce domaine.

6.1 ÉTUDES ET PUBLICATIONS

1) ABDUL GHANI

Arabic literature: uniterm indexing system for storage and retrieval / ABDUL GHANI. International Library Review; ISSN 0020-7837; GBR; 1987; Vol. 19; n°. 4; pp. 321-333.

L'auteur présente le problème posé par l'indexation en langue arabe, en prenant l'exemple de l'indexation avec des unitermes. Les difficultés d'élaboration d'un thesaurus dans le monde arabe et de normalisation du vocabulaire sont développées

2) ABED, Lamia

Génération de textes arabes dans un système de traduction anglais-arabe : Tordjmane / Lamia ABED; Ridha SALHI; Salem GHAZALI- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.91-114.

Torjmane est un prototype de traduction assistée par ordinateur développé à l'IRSIT au sein du département Arabisation. Le couple de langues traitées est : anglais-arabe. Le processus de traduction se fait en trois phases : analyse de la langue source (anglais), transfert et génération de la langue cible (arabe). L'objet de cet article est de présenter les concepts de base de TORJMANE et plus particulièrement la partie qui se charge de générer la langue arabe à partir de structures anglaises convenablement analysées par Torjmane.

3) AL BADR B.

Survey and bibliography of Arabic optical text recognition / B.AL BADR; SA MAHMOUD. Signal processing; ISSN 0165-1684; Coden SPRODR; NLD; 1995; Vol. 41; n°. 1; pp. 49-77.

Le travail de recherche sur la Reconnaissance Optique de Textes Arabes (ROTA), bien que moins avancé par rapport à d'autres langues, devient plus intensif qu'avant, et des systèmes commerciaux de ROTA deviennent disponibles. Cet article présente un aperçu et une bibliographie de la recherche sur le ROTA, couvrant toutes les publications sur le sujet auxquelles les auteurs ont eu accès. Cet article introduit le sujet plus général de la

Reconnaissance Optique de Caractères (ROC), et met l'accent sur les caractéristiques du texte arabe. Il présente également un résumé historique des systèmes de reconnaissance des textes arabes. Plus loin, ce texte fait un « état des lieux » de la recherche sur la ROTA, et énumère les spécifications des systèmes disponibles commercialement. Dans cet article, il est souligné d'abord les capacités des différents systèmes de ROTA, puis un modèle à 5 niveaux pour ces systèmes est introduit, suivi d'un classement de ce travail de recherche d'après ce modèle. Une section est consacrée à chacun des étages de ce modèle : pré-traitement, segmentation, extraction de caractéristiques, classification et post-traitement. Dans la section consacrée au pré-traitement, un accent est mis sur le traitement du texte arabe dégradé, et l'amincissement du même texte. Dans la section de segmentation, ce sont les méthodes de segmentation des textes arabes qui sont discutées suivies d'une catégorisation des méthodes selon 5 approches générales. Dans les sections d'extraction de caractéristiques et de classification, les techniques principales sont soulignées les travaux de ROTA basés sur les dites techniques sont analysées. Les auteurs discutent ensuite des approches pour le post-traitement et montrent leurs relations avec la langue arabe. Ils concluent en indiquant certains problèmes et certaines directions pour la recherche future en ROTA.

4) AL-FEDAGHI Sabah

Morphological Compression of Arabic Text / Sabah AL-FEDAGHI, Humoud B. AL-SADOON. Information Processing & Management, vol.26, n°2, 1990, pp. 303-316

La compression morphologique du texte arabe est une technique qui consiste à remplacer certains mots du texte original par leurs racines et leurs catégories morphologiques. Cette méthode est analysée à travers le développement d'une nouvelle méthode qui vise la réduction des mots arabes à leurs racines et prototypes morphologiques ainsi qu'à travers un algorithme de compression qui codifie les mots réduits dans un format à 3 octets. Cette technique est utilisée et testée par l'usage de textes multiples. Les résultats indiquent un taux de réduction entre 20% et 30% uniquement dû aux propriétés morphologiques de la langue. Cependant, 40% de réduction peut être atteint si la compression morphologique est utilisée conjointement avec une élimination des espaces du texte original.

5) AL-KHARASHI, Ibrahim

Comparing words, stems and roots as index terms for an arabic information retrieval systems / Ibrahim A.AL KHARASHI; Marta W.EVENS'- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 décembre 1993. pp.297-310

Les algorithmes du traitement des termes sur radical se basent sur l'omission des suffixes nécessaire d'une part pour réduire l'espace de stockage des listes d'autorités, et d'autre part pour élargir les éventails des réponses à des questions posées sur radical. Dans la langue arabe, l'adjonction des suffixes et des préfixes est abondamment utilisée pour donner à une racine ou à un radical quelconque une multitude de significations possibles. Un traitement en surface des affixes consiste à réduire un terme quelconque à son radical alors qu'un traitement en profondeur réduit le terme à sa racine. Ce document analyse les effets engendrés par l'usage des descripteurs de recherche sous forme de terme, racine ou radical dans un système de recherche automatisée en langue arabe. L'expérience a été menée sur un IBM/AT moyennant le logiciel de recherche documentaire Micro-AIRS et une base expérimentale de 355 enregistrements.

6) Alis ouvre les portes d' « Internet » à l'utilisateur arabe : « Le fureteur » baladeur arabisé sur le réseau. Al-Computer, communications & Electronics, pp.71-73, Janvier 1996

Il s'agit d'une synthèse des principales réalisations de la société Alis Technologies dans le domaine des systèmes multilingues (surtout arabes) pour l'accès et la diffusion de l'information scientifique et technique. Des produits comme Communiqué pour CC:Mail; Alis Multilingual Browser; Arabic Web Browser, basés entre autres sur les normes ASMO708 et Windows 1256 constituent une première en matière d'interfaces multilingues entièrement incorporés dans l'univers des grands réseaux mondiaux d'échange de l'information.

7) AMEUR, A.

L'arabe manuscrit et sa reconnaissance informatique / A. AMEU; K. ROMEO-PAKER; Y. LECOURTIER.- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 décembre 1993. pp.215-231

Cet article décrit une méthode structurale de reconnaissance de l'écriture manuscrite arabe. La lecture du texte de plusieurs lignes est faite par une caméra CCD. Le problème majeur de la lecture automatique de caractères manuscrits cursifs étant la segmentation d'un tracé en éléments constitutants, nous segmentons le texte d'abord en lignes, puis en mots et ensuite en caractères. Cette segmentation est fondée sur les propriétés contextuelles propres à l'écriture arabe. Nous nous servons de la méthode de suivi de contour et des traits de liaison entre les caractères. L'étape de la reconnaissance est étroitement liée à celle de la segmentation car elle utilise les données déjà calculées pour une pré-classification. Nous affinons la description des caractères dans un groupe plus petit par l'analyse hiérarchique qui permet le retour sur la forme pour enlever certaines ambiguïtés en cherchant les ouvertures, les occlusions, l'analyse du contour externe, les dimensions.

8) BOUDRAA, Malika

Élaboration d'une base de données arabe phonétiquement équilibrée / Malika BOUDRAA; Bachir BOUDRAA; Bernard GUERIN'- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 décembre 1993. pp.172-187

Cet article recense 20 listes de dix phrases arabes phonétiquement équilibrées. Ce corpus de références vient combler un vide bibliographique que nous avons rencontré lors de nos investigations pour effectuer différents tests d'évaluation en langue arabe (qualité de codeurs et de la parole de synthèse, évaluation de plusieurs détecteurs de mélodies, caractérisation des sons de la langue arabe, tests psycholinguistiques...). Ce corpus est basé sur les travaux de P. Combescure effectués sur la langue française. Il repose sur l'équilibre phonétique de la langue arabe donné par A.H.Moussa et M.Merayati

9) BOUYER Gérard

Transmission et réseaux de données : introduction aux éléments fondamentaux et pratiques / Gérard Bouyer.-Paris : Dunod, 1995.-450 p.

Cet ouvrage est un cours fondamental sur les techniques de transmissions et les réseaux. Dans la première partie, l'auteur s'attache aux aspects multiples des communications actuelles vues à travers les systèmes informatiques, leurs interfaces et leurs caractéristiques idéales. Le transport de l'information met en œuvre des moyens matériels développés dans la deuxième partie qui traite particulièrement des lignes, des modems et des connexions physiques entre systèmes. Dans la troisième partie, s'appuyant sur le modèle de l'ISO, il explique l'organisation des logiciels de communication au travers des primitives prévues dans la normalisation.

10) BUKHALID, Nabil

Lebanese Networking Experience / Nabil Boukhalid. In : Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tunis, 6-8 juin 1995. Cet article décrit la reconstruction en cours de l'infrastructure libanaise de télécommunications dans le cadre du Programme d'Urgence pour le Redressement National conduit simultanément avec le Programme Parallèle de la Reconstruction et du Développement. L'auteur explique l'intérêt que porte cette restructuration pour les réseaux académiques et commerciaux et la possibilité de leurs connexions internationales. Il introduit ensuite le réseau académique de recherche libanais (RARL) qui a débuté en 1994 comme étant le seul réseau non commercial dans le pays prévu pour être la plaque tournante des projets à courts et à longs termes pour mettre en place un réseau académique national de la recherche au Liban.

11) CHARTRON Ghislaine

Nouvelles problématiques pour l'IST / Ghislaine Chartron. Documentaliste Sciences de l'Information, vol.3é, n°6, 1995, pp.289-295

L'essor des nouvelles technologies, et particulièrement de réseaux comme Internet, modifie profondément les activités et produits qu'il est convenu de dénommer information scientifique et technique : le concept même d'IST s'en trouve élargi, englobant à la fois l'information liée à la pratique de la science et celle qui relève de la médiation documentaire ; pendant qu'évoluent les modes de production, diffusion, traitement et recherche de l'information. Cet article étudie ces mutations qui, brouillant les frontières des compétences, conduisent à mettre en question l'intermédiation documentaire. Et dégage pour finir les principales fonctions qui définissent aujourd'hui cette intermédiation pour l'IST dans le milieu de la recherche scientifique.

12) Client/Server architecture applications up more than 60 percent; Reengineering continues to increase and pay off for big companies,- EDI Monthly Report 8, May 1995, 3-5.

Une enquête menée par Deloitte & Touche indique que le nombre d'applications qui tournent sur des architectures du type client/serveur a augmenté de 60% et que 43% des applications industrielles sont actuellement basées sur la même architecture. 45% des 400 répondants affirment que leurs compagnies projettent investir davantage dans les systèmes client/serveur. Certaines compagnies ont investi 19% de plus qu'en 1994 pour des projets de reingénierisation. 73% parmi eux prévoient une multiplication des projets de reingénierisation et 88% expriment déjà leur pleine satisfaction des projets de reingénierisation accomplis.

13) COUSINS Shirley Anne

Enhancing subject access to OPACS : Controlled vocabulary Vs Natural languages / Shirley Anne COUSINS. Journal of Documentation, vol.48, n°3, Septembre 1992, pp.291-309

Ce document analyse un projet qui vise l'amélioration des possibilités de recherche par thème sur des catalogues publics en ligne. L'objectif de ce travail consiste à comparer les performances de recherche effectuées sur des documents indexés utilisant un vocabulaire contrôlé ou un langage naturel.

14) DESCAMPS WASSIF S.

La banque de données A'LAM de l'Institut du monde arabe / Wassif S. DESCAMPS. Documentaliste Sciences de l'Information; 1995; VOL. 32; NO. 1; PP. 39-42

Confrontée à la difficulté de gérer et diffuser un important fonds documentaire relatif au monde arabe et comprenant des ouvrages rédigés en langue arabe et en langues occidentales, la Bibliothèque de l'IMA a créé une banque de données, A'LAM, qui renferme

à la fois un fichier d'autorité et un dictionnaire bio-bibliographique des auteurs arabes. Avant de présenter ce dictionnaire et ses différentes utilisations possibles, cet article expose les choix qui ont été faits pour élaborer, en l'absence de normes, un fichier d'autorité à double vedette qui prenne en compte les spécificités du fonds et du public de l'institut du monde arabe.

15) DEVEZE Jean

L'information scientifique et technique produit scientifique ou discours sur la science. Communication et langages. N°90, 1990. pp.95-103

L'Information scientifique et technique, plus communément dite IST est devenue en peu d'années le concept à la mode, aussi prestigieux qu'incontournable pour les chercheurs de la planète. Avec, comme nous le démontre Jean DEVEZE, les risques de tout système se voulant détenteur d'une vérité à prétention universelle, une certaine intolérance entraîne l'exclusion des non-conformistes. Sait-on par exemple que certains spécialistes de l'IST prétendent chiffrer la valeur des savants en fonction d'un nombre de citations de leur nom dans les publications scientifiques ? Einstein et Freud eussent été éliminés au cours des premières années suivant leurs publications initiales. On attend le contre-pouvoir d'une certaine IST.

16) DJOUDI Mohieddine

Contribution à l'étude et à la reconnaissance automatique de la parole en arabe standard / Mohieddine DJOUDI; Jean Paul HATON (dir.) Th. doct. : Inform./Université de Nancy 1. FRA/1991/91NAN10395. 1992-11; 150 p.

La reconnaissance automatique de la parole continue pose un certain nombre de problèmes. Les caractéristiques phonétiques et linguistiques de la langue sont largement impliquées dans le processus. Ce travail consiste à présenter une contribution à la reconnaissance automatique de l'arabe standard. L'auteur a d'abord effectué une étude phonétique de la langue basée essentiellement sur l'examen de spectrogrammes de mots et de phrases en tenant compte des différents contextes de production des phénomènes. Cette étude a permis de définir les caractéristiques acoustiques des phénomènes nécessaires au système de reconnaissance. Ensuite, l'auteur a réalisé un système de décodage acoustico-phonétique, baptisé SAPHA (Système Acoustico-Phonétique de l'Arabe) qui reçoit en entrée le signal de parole d'une phrase et retourne comme résultat un treillis de phonèmes. Les principales étapes du système sont : la segmentation du signal de parole en grandes classes phonétiques (voyelles, plosives, fricatives et sonnantes); l'extraction des indices phonétiques utilisés en reconnaissance; l'étiquetage des segments utilisant un système à base de règles de production. Les méthodes utilisés ont été adaptées du système APHODEX développé pour le décodage phonétique du français. L'évaluation des performances du système a été effectuée à partir de l'étiquetage manuel des phrases phonétiquement équilibrées du corpus DJOUMA que nous avons constitué. Enfin, il a été question de développer quelques idées pour la conception d'un système de reconnaissance de phrases en arabe (MARS) intégrant le décodeur phonétique tout en soulevant les problèmes d'ordre morphologique, phonologique, syntaxique et prosodique qu'il faut résoudre

17) DJOUDI, Mohieddine

Reconnaissance de la parole en arabe moderne : le système MARS / Mohieddine DJOUDI:- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.135-143

Le système MARS (Modern Arabic Recognition System) de reconnaissance de la parole en arabe moderne. Le système a pour but la compréhension de phrases dans un contexte

multilocuteur. Sa mise en œuvre tient compte des particularités de la langue aux niveaux phonétique et linguistique. Les autres thèmes abordés : les aspects de la morphologie de la langue dans le traitement linguistique, la structure syntaxico-sémantique des phrases, l'importance de la prosodie dans un système de compréhension de l'arabe parlé.

18) DUCASSE Roland

Modèle Client-Serveur et systèmes d'information distribués : éléments pour une reingénierie des services d'information scientifique et technique. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. pp.133-138.

Cet article confirme les avantages techniques et informationnels des architectures client-serveur et des systèmes distribués. À part le besoin de revoir en toute sa profondeur la notion de l'intermédiation et du rôle de l'utilisateur final dans le processus informationnel et communicationnel, l'auteur évoque l'importance des normes dans la transparence de l'hétérogénéité des systèmes. Désintermédiation de l'accès à l'IST, décentralisation des sites de traitement des données, banalisation des procédures de communication.. constituent quelques-uns des atouts qui convergent vers une réingénierie des services d'information scientifique et technique du mode centralisé au mode distribué, de l'architecture terminal-serveur au modèle client-serveur.

19) EL ALAOUI Faris M.

Alexie sans agraphie en langue arabe : étude neurolinguistique et IRM / Faris M. EL ALAOUI; BENBELAID F; ALAOUI C; TAHIRI L; JIDDANE M; AMARTI A; CHKILI T. Revue neurologique : (Paris); 1994; vol. 150, n°11, pp. 771-775.

Une femme de 33 ans présentait une alexie sans agraphie, une anomie des couleurs, une hémianopsie latérale homonyme droite, une aphasie amnésique et une amnésie verbale. L'IRM cérébrale montra des lésions du splénium du corps calleux, du forceps major, des radiations optiques et du lobe temporal antérieur gauches. La conservation de la compréhension écrite et la récupération complète des troubles de la lecture malgré la persistance des lésions plaident en faveur d'une participation active de l'hémisphère droit à la lecture; cette capacité de l'hémisphère droit serait due aux particularités de l'écriture arabe

20) ELLIS David

The physical and cognitive paradigms in information retrieval research / David ELLIS. Journal of Documentation, vol.48, n°1, Mars 1992, pp.45-64

Les caractéristiques de deux paradigmes agissant dans le domaine de la recherche de l'information sont discutées dans ce papier : le paradigme physique et le paradigme cognitif du point de vue de leurs origines, natures et rôles respectifs. Il a été démontré que bien que la majeure partie des travaux effectués dans le domaine de la recherche de l'information relèvent des domaines des paradigmes physiques et cognitifs, aucun d'eux ne prétend constituer à lui seul une science établie.

21) FARROW John F.

A cognitive process model of document indexing / John F.FARROW. Journal of Documentation, vol.47, n°2, Juin 1991, pp.149-146

La classification, l'indexation et la condensation peuvent être considérés comme des opérations pour résumer le contenu des documents. Un modèle de compréhension de texte par les indexeurs est présenté basé sur un principe de description de tâches qui considère que la compréhension de textes pour l'indexation diffère d'une lecture courante. Ceci est dû entre autres à : les contraintes de temps de lecture qui fait que le texte est plutôt scanné

rapidement pour rassembler un minimum d'indices de compréhension; la compréhension étant plutôt orientée tâche (et non apprentissage), elle est immédiatement suivie par la production d'un résumé ou d'un index; le réflexe automatique des indexeurs expérimentés travaillant dans une catégorie de texte limitée. L'évidence de l'interconnexion entre l'analyse conceptuelle et l'analyse perceptible d'un texte dans un contexte de lecture rapide est remise en cause. L'attribution de ressources mentales pour l'analyse de texte est discutée suivie par un modèle de traitement cognitif est proposé.

22) GOODMAN S.E.

Computing in the Middle East / S.E. GOODMAN, J.D. GREEN. Communication of the ACM, vol.35, n°8, Août 1992, pp.21-2

Généralement, les couvertures médiatiques étrangères au Moyen Orient, miroitent une image erronée de peuples qui refusent le modernisme et l'acceptation des technologies provenant de l'Occident et de vivre sur des valeurs qui remontent à des époques historiques lointaines. Pour l'auteur cette impression d'une technophobie régionale est très simplifiée et même exagérée. Comme les exemples qu'ils cite le laissent démontrer, il y a eu en fait une implication importante de nouvelles technologies dans la région dont certaines acquièrent des caractéristiques indigènes très particulières. Déjà, comme c'est démontré, des circonstances économiques, politiques et culturelles sont d'une telle ampleur que l'acceptation de ces technologies reste assez ambiguë.

23) HASSANEIN, Oussama

The global information superhighway : issues for the arab region / Oussama HASSANEIN. In : Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tunis, 6-8 juin 1995.

Durant les 4 premiers mois de l'année 1995, l'industrie de Télécommunication américaine a continué à pas de géants son progrès vers une meilleure qualité des services multimédia offerts sur des principes de la concurrence des prix, de la conception orientée usager et de l'ubiquité. Au moment où le Congrès des États Unis discutait une nouvelle loi sur la communication afin d'encourager la concurrence, les services PTT ainsi que d'autres pourvoyeurs de services de télécommunication subissaient les aléas des services d'interconnectivité d'un marché vertical. Dans cette présentation, l'auteur expose les tendances techniques, réglementaires et économiques que prendront ces services à long terme comme exprimés par les services PTT ainsi que les commerçants en télécommunications eux mêmes.

24) JAMBI, Kamel M.

An approach for segmenting handwritten arabic words / Kamel M. JAMBI- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.233-243

Ce document expose une nouvelle approche de segmentation du texte arabe manuscrit. Après une brève analyse des travaux déjà accomplis dans ce sens, cette étude décrit en détail le processus en question. Un taux de segmentation de 95% est obtenu et quelques interprétations des résultats sont avancées.

25) JAVED Mostafa

Design and analysis of the human-computer interface for an image database (Database) / Mostafa, Javed. Thesis (PH.D.)--The University of Texas at Austin, 1994, 124p..-(Source : Dissertation Abstracts International. Volume: 55-10, Section: A, page: 3023.)

Concevoir des systèmes de recherche d'images sur des bases de données est un problème aussi complexe que celui de la conception des systèmes sonores pour l'aide à la recherche. L'interface Homme-Machine est la composante principale capable de faciliter l'opération de recherche sur une base de données. La recherche de sources graphiques pourrait donc s'améliorer grâce à une amélioration dans la conception des interfaces Homme-Machine. Une IHM a été créée avec deux moyens de recherche d'images : visuel et verbal. Un environnement orienté objet et une architecture client/serveur ont servi pour développer les dites interfaces. Un ensemble d'images animées numériques extraites de films populaires a été utilisé comme échantillon des tests accomplis. Tout en utilisant les deux techniques manuelles et automatisées, une procédure pratique d'indexation d'images a été mise en place. L'index comprenait des informations verbales et visuelles entièrement intégrées dans l'IHM visuelle et verbale comme aide de recherche en ligne. L'efficacité des deux IHM a été comparée grâce à des données quantitatives et qualitatives collectées par des opérations de recherches couvrant 18 utilisateurs. Aucune différence fondamentale n'a été observée entre les deux IHM au niveau des variables concernant l'accomplissement des recherches, les erreurs, les demandes d'aide en ligne et l'optimisation des recherches. Il est à signaler cependant que dans les deux cas, les taux de l'accomplissement des recherches dépassaient les 80%.

26) KURDI, Mohsen Bassam

Multifont arabic/latin optical character recognition system / Mohsen Bassam KURDY; Ammar JOUKHADAR; Ahmad WABBI- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.245-256.

La reconnaissance optique des caractères peut largement contribuer à l'avancement des procédés d'automatisation, à la perfection des interfaces Homme-Machine et à la création de nouvelles applications dans le domaine de la Bureautique, de la gestion des données et surtout à la consolidation des systèmes d'information (i.e. les systèmes d'archivage optique). Ce document présente un système de reconnaissance multifontes du texte imprimé arabe/latin qui inclut une approche morphologique mathématique de la reconnaissance des caractères. Un algorithme de reconstitution d'image est appliqué pour segmenter les caractères arabes rattachés et superposés. Étant donné le nombre important de formes à identifier, cette nouvelle approche utilise un critère de reconnaissance à deux niveaux d'identification des formes : la description et l'interprétation. Le niveau d'interprétation est accompli grâce à des paramètres non métriques. Son objectif est de classer les caractères en plusieurs catégories qui serviront d'éléments d'analyse pour le deuxième niveau. L'interprétation et la distinction entre ces groupes de caractères sont entièrement heuristiques. Plusieurs de ces cas sont illustrés et discutés dans cette étude.

27) LAINE-CRUZEL Sylvie

Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs / Sylvie LAINE-CRUZEL. Documentaliste Sciences de l'Information, vol.31, n°3, 1994, pp. 143-147

L'idée dominante, au sujet de l'indexation est que le sens d'un texte est contenu dans le document : pour aboutir à une représentation du sens, il faut analyser aussi finement que possible la structure apparente du texte, c'est à dire identifier et caractériser, par des outils linguistiques et/ou statistiques, la forme que prend ce texte, reflet exact du sens que l'auteur lui a donné. Sylvie Lainé-Cruzel présente dans cette étude d'autres modèles - encore théoriques - qui s'appuient sur l'idée que le sens est au contraire construit par le lecteur, et qu'il est donc différent d'un lecteur à un autre : dans un système d'information, une bonne

représentation du contenu d'un document passe donc par la prise en compte de certaines caractéristiques de son futur utilisateur. devant la multiplication des données gérées et la variété des usages qui en sont faits, une solution d'avenir est sans doute, comme celle qui est proposée ici, dans l'association de ces deux approches.

28) MILOUDI, M.T.

Approche système expert pour la traduction automatique réversible du Braille intégral ou abrégé vers l'arabe / M.T. MILOUDI; Z. ZMERLI- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.371-381

La tendance actuelle de l'automatisation signifie que les déficients visuels ont besoin d'un meilleurs accès à l'information. L'édition en braille doit être disponible rapidement et la communication écrite entre les non-voyants doit se normaliser. Les traductions Braille-noir et noir-Braille ont été axés sur une approche système expert ou nous avons dissocié l'aspect élaboration des règles de traduction de l'aspect moteur de traduction. La base de règles concernant la traduction Braille-noir a été soigneusement étudiée dans ses aspects de simplicité d'élaboration et de test pour les performances du traducteur et de mise à jour par une personne non-informaticienne. La traduction noir-Braille a été plus rapidement mise en oeuvre. Le caractère déterministe des règles élaborées lors de la traduction Braille-noir a rendu la base de règle complètement réversible. Le moteur quand à lui a été réaménagé en fonction des particularités inhérentes à cette traduction.

29) MNIF, Mohamed

Arabic/Latin Bilingual Videotex System / Mohamed MNIF. Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tunis, 6-8 juin 1995.

Ce projet vise le développement d'un système vidéotex bilingue (arabe/latin) et l'élaboration de ses spécificités techniques conformément au contexte culturel arabe. Deux recommandations internationales sont prises en considération : CCITT T 101 qui concerne l'interconnexion internationale des services vidéotex ; CCITT T 52 qui définit les caractères non latins pour les services télématiques.

30) MONEIMNE Walid

AO vers l'arabe : spécification d'une génération standard de l'arabe : réalisation d'un prototype anglais-arabe à partir d'un analyseur existant / Walid MONEIMNE; Christian BOITET (Dir. thèse); Th. doct. : Inform./Grenoble 1/1989/89 GRE1 0061; 1989; 259 p.

La thèse se situe dans un ensemble d'études linguistiques et informatiques visant à voir dans quelle mesure la méthode linguistique et les outils informatiques du GETA s'appliquent à la langue arabe. Après une brève synthèse des études antérieures en traitement automatique des langues naturelles appliqué à l'arabe, l'auteur rappelle les fondements de la traduction assistée par ordinateur ainsi que les aspects théoriques et méthodologiques les plus marquants de l'approche suivie au GETA. Cette approche consiste essentiellement à travailler simultanément à différents niveaux d'interprétation de l'arabe. L'auteur rappelle également les fondements de la traduction assistée par ordinateur ainsi que les aspects théoriques et méthodologiques les plus marquants de l'approche suivie au GETA. Cette approche consiste essentiellement à travailler simultanément à différents niveaux d'interprétation de description linguistique en programmant dans des langages spécialisés pour la programmation linguistique (LSPLs). En est montré en détail comment a été spécifié et installé un logiciel de traduction de l'anglais vers l'arabe; est spécifiée une grammaire statique d'un sous-ensemble de la langue arabe à l'aide du formalisme des grammaires

statiques (correspondances arbre-chaîne). Pour spécifier la grammaire du transfert, l'auteur propose une généralisation de ce formalisme aux correspondances arbre-arbre. Tout au long de la thèse, sont proposés des exemples et des traces d'exécution permettant de suivre les principes de la méthode utilisée.

31) NAZIM Ali S.

CD-ROM databases as an alternative means to online information: the experience of a university library in developing countries / Ali S. NAZIM. International online information meeting. 12/1988-12-06/London (Londres); GBR; Oxford: Learned Information Europe, 1988. Vol. 2; pp. 569-573.

Cette expérience est décrite d'une part par son contexte: bibliothèques universitaire n'ayant pas encore pu résoudre les problèmes de sa propre automatisation (Langue arabe/non-arabe, coût...), et suivant les trois ans: sélection du matériel, sélection des bases de données, mise en place des services d'accès aux CD-ROM. À long terme, cette opération a une conséquence sur la nécessité du développement des collections

32) NOUH A.

Recognition of constrained handprinted Arabic Character / A.Nouh- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.201-214

Ce document présente un algorithme de reconnaissance du caractère comprimé de l'arabe manuscrit. La technique fait recours à un tableau de codage proposé et une procédure spéciale de normalisation antécédente à l'étape de reconnaissance. Cet algorithme part de l'idée que les largeurs « tentaculaires » du caractère peuvent varier à volonté donc sont passibles d'être repérées dans les limites des références des caractères. Il est donc possible de les utiliser comme des « templates » des bordures internes ou externes. Une expérience menée sur la base de cet algorithme dans le but de reconnaître 56 caractères manuscrits comprimés a donné un taux de réussite de 95.8%.

33) POLITY Yolla

Évaluation des modes de recherche en langage naturel / Yolla POLITY. Documentaliste Sciences de l'Information, vol.31, n°3, 1994, pp.136-142

Le problème de l'interrogation en langage naturel des banques de données documentaires est ici abordé dans le cadre général de l'ergonomie des interfaces hommes-machine pour les systèmes d'information. Cette approche conduit Yolla Polity à s'interroger sur la place respective des différents modes de recherche d'information en langage naturel en tenant compte de deux paramètres principaux : les types d'utilisateur et les types d'information à rechercher. Les travaux entrepris dans ce domaine partent de l'étude de la tâche de recherche d'information et se dirigent ensuite dans deux directions : dépasser le stade actuel d'une interaction limitée à des questions-réponses indépendantes pour arriver à gérer un dialogue Homme-Machine ; mettre à la disposition de l'utilisateur un accès multimodal dans lequel le langage naturel sera un mode d'interaction parmi d'autres. Ces travaux exploitent les progrès de l'informatique et de l'ingénierie linguistique, ainsi que les apports de la psychologie cognitive.

34) RAJOUANI, A.

Synthèse de l'arabe à partir du texte : développement et réalisation d'un système de haute qualité / A. Rajouani; M. Ouadou; M. Najim; M. Zyoute' In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al

Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.117-133

Une partie de cet article concerne le développement et la mise au point d'un système de synthèse par règles de langue arabe qui permet de transformer tout texte voyellé en un message vocal intelligible et relativement naturel. Le système est modulaire et contient 5 modules dont nous présentons pour chacun l'entrée, la sortie, la fonction et les caractéristiques générales ainsi qu'une évolution sommaire. La deuxième partie décrit la réalisation et l'implantation en temps réel du synthétiseur à formants de Klatt série et parallèle sur un processeur de signal.

35) SERRES Alexandre

Hypertexte : une histoire à revisiter / Alexandre SERRES. Documentaliste Sciences de l'Information, vol.32, n°2, 1995, pp.71-83

S'appuyant notamment sur des travaux de Buckland, Nyce et Kahn, Alexandre Serres retrace ici la genèse et les premiers développements de la notion d'hypertexte : un mode de lecture écriture non linéaire qui s'est concrétisé à travers des techniques très différentes sans pouvoir être réduit à l'une d'elles. L'évolution de cette notion hybride se situe à la croisée de plusieurs histoires, celles de l'informatique, de la documentation, de l'imprimerie, de l'utopie du savoir universel, des technologies intellectuelles.. Cette recherche conduit à relativiser, sans le diminuer, le rôle éminent de Bush, et à admettre une multiplicité d'origines simultanées de l'hypertexte. Et à rendre à Paul Otlet la part qui lui revient dans la préfiguration de cette idée : si Bush en a exposé les principes qu'il a commencé à concrétiser en « inventant » le Memex et le sélecteur rapide de microfilms, les conceptions d'Otlet, sans doute moins précises, apparaissent beaucoup plus ouvertes et visionnaires.

36) VICKERY Brian

An application of language processing for a search interface / Brian VICKERY, Alina VICKERY. Journal of Documentation, vol.48, n°3, septembre 1992, pp.255-275

Le document décrit une technique développée par Tome Associates pour traiter des requêtes exprimées en langage naturel en équations de recherche transmissibles vers des systèmes de bases de données textuelles en ligne. Les problèmes évoqués dans ce document concernent : l'identification des mots, la manipulation de mots non identifiés, les contenus et les structures des dictionnaires systèmes, l'usage des catégories sémantiques et des classifications, la distinction entre des mots à plusieurs sens, la radicalisation et la troncature, les mots composés et la relation entre les termes de recherche.

37) VICKERY Brian

Online search interface design / Brian VICKERY, Alina VICKERY. Journal of Documentation, vol.49, n°2, juin 1993, pp.103-187

Il existe une quantité énorme de données stockées dans des bases et banques de données accessibles en ligne constituées de grands fichiers en texte récupérable par les techniques booléennes de recherche. Il est cependant communément admis que l'usage de ces sources d'information est beaucoup moins intense que prévu et que parmi les raisons de cette restriction d'usage est le fait que l'utilisateur potentiel trouve assez complexe le processus d'identification de la base de donnée à interroger, l'usage des différentes commandes de dialogue avec les serveurs et de construire l'équation booléenne de recherche pertinente à son besoin informationnel. Ceci a engendré un effort considérable d'analyse et de recherche pour aboutir à la conception d'interfaces de recherche capable d'aider les utilisateurs profanes à tirer un maximum de profit de ces mêmes bases de données.

L'objectif de ce document est de revoir les aspects de la conception de ces interfaces pour indiquer le nécessaire devant être accompli pour garantir un maximum d'aides et de convivialité aux utilisateurs inexpérimentés ; pour mettre en relief le savoir-faire qui doit accompagner la conception de cet aspect d'assistance ; pour décrire certaines solutions utilisées dans des interfaces expérimentales et opérationnelles ; et pour discuter de certains problèmes rencontrés.

6.2 DES SITES WEB

L'internationalisation (i18n)^[77] et la localisation (l10n) des programmes et des données dans les domaines de la Bureautique et d'Internet gagnent progressivement en qualité grâce à la contribution de groupes de chercheurs dans ce domaine particulier. Ainsi l'Internet Multilingue se développe sur deux courants parallèles : le développement multilingue du contenu et l'évolution des outils logiciels « internationalisés » et « localisés » pour un grand nombre de langues non latines (i.e. arabe, tchèque, grec, hébreu, hongrois, polonais, russe, chinois...).

Le multilinguisme arabe/latin, bien que retardataire par rapport aux réalisations accomplies par d'autres langues non latines, comme c'est le cas des langues asiatiques (Chinois, Japonais et Coréen surtout), occupe progressivement sa place dans la littérature mondiale grâce aux différents sites Internet qui proposent des solutions multilingues. Ces initiatives ressortent tout particulièrement du domaine commercial. La contribution de secteur académique et de la recherche est également notoire.

La liste des services que nous présentons ci-après constitue selon notre investigation, les références les plus appropriées pour mettre en place un système de veille technologique en la matière. Nous estimons, compte tenu de la nature de leurs produits multilingues, qu'elles resteront pour longtemps des sources d'information importantes sur ce domaine de recherche et plus particulièrement arabe-latin.



Alis Technologies : (<http://www.Alis.com>)

Depuis sa création en 1981, la société Alis Technologies travaille sur les solutions informatiques de traitement multilingues. Son domaine d'expertise couvre les langues de l'Europe de l'Est et de l'Ouest, du Moyen-Orient ainsi que quelques langues asiatiques.

Le champ d'action d'Alis Technologies couvre les domaines des technologies linguistiques essentiellement avec deux solutions multilingues :

- [Batam](http://montreal.alis.com/castil/batam.fr.html) : une bibliothèque d'internationalisation qui offre un support linguistique varié permettant aux applications multilingues d'afficher, d'imprimer et d'effectuer la saisie de texte dans plus de 90 langues sans contraintes de la plate-forme Windows de travail. (<http://montreal.alis.com/castil/batam.fr.html>).
- [Flores](#) : une trousse de composants internationalisés qui peuvent s'intégrer à n'importe quelle application, qu'elle soit écrite pour Windows ou non pour

[77] Symbole utilisé dans le monde du développement des systèmes d'information multilingue comme acronyme ou abréviation du terme Internationalisation : i + 18 lettres + n.

accomplir des fonctionnalités comme le bris de lignes, le tri des chaînes de caractères, le rendu visuel multilingue (affichage) et les filtres de conversions de jeux de caractères (également appelés transcodage). (<http://montreal.alis.com/castil/flores.fr.html>)

Le plus important produit de la société Alis Technologies est sans doute le kit multilingue **Tango** qui constitue une suite d'outils Internet multilingue considérée actuellement à la pointe des solutions de l'internationalisation d'Internet. Cette suite comprend les trois outils de bases :

1. **Le navigateur Tango** : un butineur dans plus de 90 langues ;
2. **Tango Courrier** : un programme de courrier électronique dans plus de 90 langues ;
3. **Tango Creator** : un éditeur de pages Web multilingues.

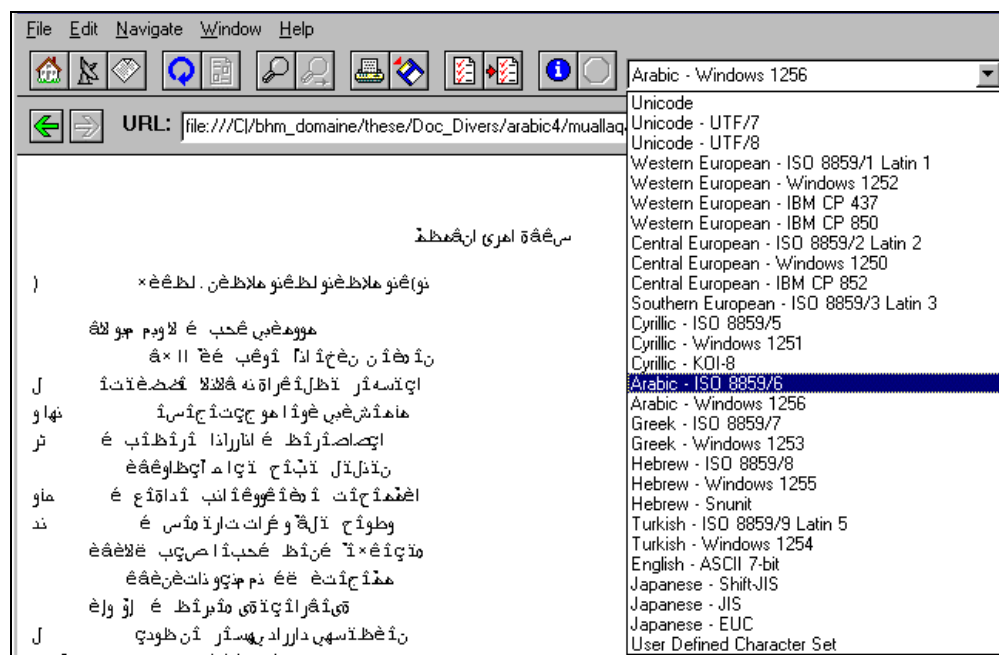


Accentsoft : (<http://www.accentsoft.com>)

Accent Software offre en plus de leurs solutions de systèmes d'aide à la traduction, une suite de 4 programmes pour la navigation et l'édition sur Internet.

Internet With an Accent inclut un navigateur Web, un éditeur WYSIWIG de pages HTML, un éditeur de messagerie et un outil public de visualisation (viewer) de tous documents créés par la suite concernée.

Les solutions Accent Software permettent de lire et d'écrire en sur Internet en 30 langues y compris les langues arabe, bulgare, croate, finlandaise, roumaine, etc. les menus systèmes peuvent s'adapter à une quinzaine de langues différentes.





Sakhr : (<http://www.sakhr.com>)

Les programmes de la société Sakhr sont répertoriés par catégories classées par domaines de spécialité qui varient entre les programmes Internet, les programmes éducatifs et de familles, les programmes de Bureautique ... tous ayant la langue arabe comme langue de base.

Dans le contexte Internet, les solutions Sakhr sont également orientée vers le concept de suite logicielle qui vise la navigation et la recherche, le courrier électronique ainsi que l'édition et la création des pages HTML en mode multilingue.

Parmi les produits Internet Multilingues de Sakhr :

[Sindabad](#) : est un plug-in multilingue conçu comme extension au navigateur Netscape communicator pour répondre aux exigences de la langue arabe tant au niveau de la navigation que de la messagerie, de l'édition, de la technologie « Push », des conférences électroniques...

(<http://www.sakhrsoft.com/lproducts/linter/sind4.htm>)

[Nasher Net](#) : est un outil WYSIWYG d'édition électronique de troisième génération pour la création des pages Web en langue arabe. Nasher Net tourne sur toutes les versions de versions de Microsoft Windows 95 dans n'importe quelle langue avec des possibilités d'extraction automatique des mots-clés qu'il insère dans les champs métadonnées des pages Web.

(<http://www.sakhrsoft.com/lproducts/linter/nasher.htm>)

[AlDalil](#) : C'est le répertoire Internet des ressources d'information en langue arabe (exclusive !) accessibles par les deux méthodes courantes d'accès :

- Recherche sur répertoires pré-indexés en structure arborescente
- Recherche sur texte intégral grâce à un robot indexeur et un moteur de recherche.

(<http://www.sakhrsoft.com/lproducts/linter/dalil.htm>)

[Al Idrisi Internet](#) : un moteur de recherche de documents multilingues sur Internet. La langue arabe y est une langue d'interface et de recherche d'information.

À l'instar de Yahoo qui fait appel au moteur AltaVista pour les opérations de recherche élémentaires, Al Idrisi^[78] est le moteur d'appui de Al Dalil.

(<http://www.alidrisi.com/alidrisi.htm>)



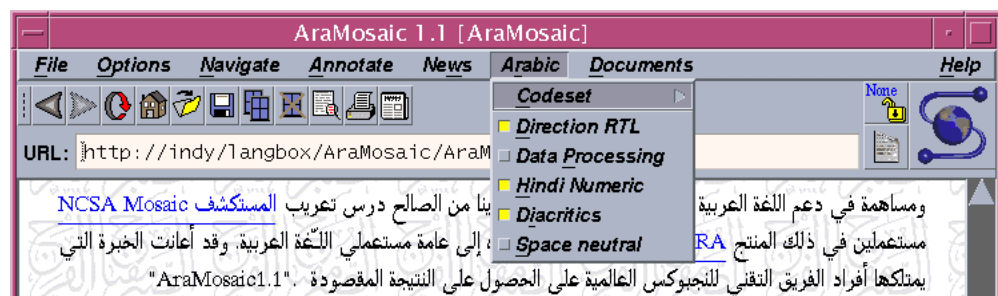
AraMosaic : (<http://www.langbox.com/AraMosaic/index.html>) ¹

C'est la version aménagée du navigateur standard Mosaic (version 2.7b4) de la NCSA (National Center for Supercomputing Applications) conçu et distribué par l'université de l'Illinois aux États Unis d'Amérique pour tourner sur la plateforme Unix/X11. AraMosaic est l'alternative à son prédécesseur PMosaic avec

[78] IDRISI = Intelligent Document Retrieval and Information Search for the Internet

support trilingue anglais, arabe et persan. Étant donné que PMosaic ne fonctionne pas à base de la norme de codification des caractères ISO 8859-6 (§ III.3.3.1) mais plutôt à base de la norme persane (farisi) ISIRI 3342, la société LangBox International a mis au point l'alternative AraMosaic sur la base du code ISO 8859-6 pour la communauté des utilisateurs d'Unix en langue arabe moyennant son outil de développement XLANGBOX-ARA.

Sous cette base, AraMosaic fonctionne conformément au protocole HTTP standard pour l'échange des documents anglais/arabe en mode HTML avec des perspectives de développement futurs pour pallier certaines lacunes au niveau de la portabilité multiplates-formes, de la gestion de la bidirectionnalité, de l'analyse de contexte, de l'impression en postscript ...



Video On Line : (<http://www.vol.it>).

VOL est en principe un service d'accès à Internet basé en Italie. Mais à part le simple service de relais à Internet, VOL offre un service d'accès à des ressources d'information multilingues sur la base d'une répartition géographique et non linguistique. Chaque document détient une équivalence linguistique d'un pays quelconque. Quarante pays sont représentés dans le choix linguistique des services VOL. À titre d'exemple, trois versions anglaises différentes (britannique, américaine et canadienne) disposent de trois pointeurs différents. VOL offre un navigateur multilingue public nommé « Tiber » qui fonctionne en 18 langues différentes.



Unicode (<http://www.unicode.org>)

The official Web site for the Unicode consortium provides all the updates on Unicode Standard versions, international Unicode conferences, and information about language encodings.



W3C (<http://www.w3.org/pub/WWW/International/Overview.html>)

Ce site du World Wide Web Consortium (W3C) recense un grand nombre de ressources concernant l'internationalisation et la localisation des ressources Internet : Internationalisation de HTML et HTTP, les URLs multilingues, l'encodage des jeux des caractères, les feuilles de styles pour les langues non-latines.



ISO (<http://www.iso.org>)

L'organisation internationale de normalisation (ISO) est une fédération mondiale d'organismes nationaux de normalisation de quelques 100 pays, à raison d'un organisme par pays.

L'ISO est une organisation non gouvernementale, créée en 1947. Elle a pour mission de favoriser le développement de la normalisation et des activités connexes dans le monde, en vue de faciliter entre les nations les échanges de biens et de services et de développer la coopération dans les domaines intellectuel, scientifique, technique et économique.

Le champ d'action de l'ISO ne se limite pas à un secteur particulier. Il couvre tous les domaines de la normalisation, à l'exception de l'ingénierie électrique et électronique, qui sont du ressort de la CEI. Les travaux dans le domaine des technologies de l'information sont menés par un comité technique mixte ISO/CEI (JTC 1).



Localization Industry Standards Association (LISA)
(<http://www.lisa.unige.ch/>)

LISA est un groupe professionnel de support pour le développement des standards de localisation de logiciels.

CHAPITRE DEUXIÈME

LES ASPECTS DE LA COMMUNICABILITÉ DE L'INFORMATION AU DIAPASON DES DISCIPLINES CONNEXES

Avant d'aborder la dimension des interfaces de communication dans des systèmes de communication multilingues, nous avons jugé important de faire précéder cette partie dédiée à l'ergonomie des Interfaces Homme Machine par une analyse structurelle du fonctionnement interne des systèmes d'information automatisés comme annoncée dans la problématique du sujet. Notre but est de lever le rideau sur le fonctionnement interne de ces systèmes et les particularités de ses méthodes pour le traitement des langues afin de préciser les possibilités que nous jugerons les plus adéquates. Nous nous positionnons comme évaluateur des interfaces Homme-Machines multilingues, au point de contact entre les deux protagonistes avec un regard analytique sur le fonctionnement de l'un et l'attitude de l'autre. D'un côté comme de l'autre, nous sommes contraints d'aller plus en amont des facteurs parallèles qui agissent directement ou non sur la structure, la qualité et le fonctionnement des deux pôles d'interfaçage : la machine et l'homme.

Cette perspective ne saurait, à notre sens, échapper en début de parcours à une exposition du contexte général des grands aléas de l'histoire des phénomènes d'information et de communication et des avancées conceptuelles et surtout technologiques qui l'ont toujours marquée depuis leur origine du moins moderne et contemporaine. Ceci nous oblige à revoir les évolutions qu'ont connues ces deux concepts depuis le temps des pères fondateurs.

Ce qui est évident, a priori, c'est que l'information et la communication ont largement évolué, dans une perspective moderniste, au-delà des définitions traditionalistes d'autrefois. Avec la nouvelle théorie de l'information de Shannon, elles allaient être marquée par les évolutions de domaines connexes déterminants pour la pensée collective de chaque époque et la perception des phénomènes constitutifs de la théorie shannonienne basée sur la trilogie « émetteur, message, récepteur » ([Fig.2](#)). Les sciences linguistiques et la sémiologie, le traitement du signal et le codage binaire, la numérisation et l'algorithmique sont autant de phénomènes qui auront à notre sens, une influence considérable sur les théories et les systèmes d'information et de communication. Un aperçu des divers aspects de l'intervention de ces facteurs nous aidera à positionner la problématique de notre étude dans un contexte catalysant.

1. LES THEORIES LINGUISTIQUES ET LA SEMIOLOGIE DU LANGAGE : DE L'ABSTRACTION PLATONICIENNE AU STRUCTURALISME SAUSSURIEN

Les théories linguistiques, même sous d'autres appellations, ont toujours constitué un champ d'analyse et de réflexion de la dynamique des rapports de l'homme avec son milieu naturel. Or, ce qui nous intéresse dans l'étude actuelle, c'est l'aspect représentatif, symbolique ou simplement sémiotique de la langue, et l'évolution qu'a connu la définition du langage et plus tard de la linguistique depuis les théories « abstractionnistes » de l'antiquité et de l'ère médiévale jusqu'aux tendances empiriques et structurales de la linguistique moderne. À travers ces périples dans l'évolution de l'étude des langues, la polémique a particulièrement pris la forme d'un débat autour des mêmes dédoublements et des mêmes antinomies sur lesquels repose la dialectique linguistique de tous les temps : langue et parole, identité linguistique et système, synchronie et

diachronie, signes et sens, signifiant et signifié, linéarité et hiérarchie, arbitraire et conventionnalisme, paradigme et syntagme du mécanisme de langue.

1.1 À LA RECHERCHE D'UNE DÉFINITION DE LA LANGUE ET DU LANGAGE

Cette multitude d'agréats attribués à la langue et au langage allait se refléter en effet sur les travaux et les définitions formulées dans les écrits des grands spécialistes en la matière à travers les temps.

1.1.1 Quelques définitions du XVIII^e siècle

- ❑ Le *Dictionnaire de l'Académie française* (1694) précise : « *Langage : idiome, langue que parle une nation* » ; « *Langue : idiome, termes et façon de parler dont se sert une nation* ».
- ❑ Antoine Furetière (1619-1688), dans son *Dictionnaire* (édition de 1704), identifie les deux termes de langue et de langage de manière presque similaire mais rajoute cependant : « *Langage : suite de paroles dont chaque peuple est convenu ; langue : en usage chez une nation pour expliquer les uns aux autres ce que l'on pense (les étrangers n'entendent point notre langage).* »
- ❑ Selon Frain du Tremblay (1713), le langage est « *une suite ou amas de certains sons articulés propres à s'unir ensemble, dont se sert un peuple pour signifier les choses et pour se communiquer ses pensées : mais qui sont indifférents par eux-mêmes à signifier une chose plutôt qu'une autre* ».
- ❑ Dans l'*Encyclopédie* de Diderot^[1] (1755) il est mentionné : « *Une langue est la totalité des usages propres à une nation pour exprimer les pensées par la voix* ».
- ❑ Le *Grand Dictionnaire universel du XIX^e siècle* (Larousse) précise : « Une langue est l'idiome d'une nation. »
- ❑ « *Parler, c'est expliquer sa pensée par des signes que les hommes ont inventés à dessein* », disait la *Grammaire générale et raisonnée de Port-Royal*^[2] pour confirmer la tendance aristotélicienne du langage comme outil de la pensée. Cette tradition d'Aristote, en effet, définissait la langue autant par son but (le fait d'exprimer la pensée ou de la communiquer) que par son moyen (un système de signes).
- ❑ Des linguistes comme J. Frain de Tremblay (1713) et Antoine Furetière (1619-1688) ont, en effet largement contribué à la confirmation de cette théorie d'Aristote attribuée également à Descartes concernant l'arbitraire dans l'assignation des signes pour représenter le langage. « *Le principe de l'arbitraire n'est contesté par personne* »^[3] confire plus tard Ferdinand De Saussure.

[1] Œuvre encyclopédique de D. Diderot éditée par J. Le Rond D'Alembert et intitulée *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, 17 vol. De texte et 11 vol. De planches. Paris, 1751-1772, rééd. Fac sim., 35 vol., Stuttgart, 1966-1967.

[2] PORT-ROYAL (LOGIQUE DE). On désigne habituellement sous le nom de *Logique de Port-Royal* un ouvrage paru en 1662 et intitulé *La Logique ou l'Art de penser*, qui fut élaboré dans les Petites Écoles de Port-Royal par Antoine Arnault et Pierre Nicole. Destiné à être un manuel abrégé, permettant aux non-spécialistes « d'apprendre en quatre ou cinq jours tout ce qu'il y a d'utile dans la logique », il représente, en outre, une synthèse des principaux courants de pensée du XVII^e siècle (à l'exclusion du scepticisme et du matérialisme).

[3] Ferdinand DE SAUSSURE. *Cours de linguistique générale*. Edition critique préparée par Tullio De Mauro. Payot. Paris : 1986. p.100.

- ❑ D'Alembert (1717-1783), dans le « Discours préliminaire », emploie la même idée d'une « collection assez bizarre de signes de toutes espèces »^[4] pour caractériser l'origine des langues, c'est-à-dire le stade où il n'existait pas encore d'*usages* établis.

Tel est l'état des traits assez variables que les auteurs différents retiennent pour caractériser le concepts de langue et de langage, conformément à leurs référents sociolinguistiques, leurs réflexions, leurs orientations philosophiques.

1.1.2 Quelques définitions du XX^e siècle

Depuis le début du XX^e siècle de nouvelles définitions des langues ont émergé pour constituer une nouvelle approche d'analyse et une nouvelle démarche dans le traitement des langues. La nouvelle dimension du code linguistique ou « signe linguistique » émerge dans sa plénitude dans les définitions modernes pour concevoir la langue comme un système de signes à partir desquels un moyen d'expression et un instrument de communication seraient établis.

- ❑ Pour Ferdinand de Saussure (1857-1913), la définition de langue comme énoncée dans le *Cours de Linguistique Générale*^[5] (Paris, 1915) est « un système de signes distincts, correspondant à des idées distinctes ».
- ❑ Dans son ouvrage *Language* (New York, 1921), Edward Sapir (1884-1939) évoque la définition du langage comme un « moyen de communication ... par l'intermédiaire d'un système de symboles ».
- ❑ Pour le *Vocabulaire technique et critique de la philosophie* de Lalande (1867-1963) (1926), le langage est « au sens le plus large, tout système de signes pouvant servir de moyen de communication ».
- ❑ Otto Jespersen^[6] cité dans l'*Encyclopædia Britannica* (1932) identifie la langue comme étant « n'importe quel moyen de communication entre les êtres vivants ».
- ❑ Dans le *Lexique de la terminologie linguistique* (3^e éd., 1951) de J. Marouzeau^[7], c'est « tout système de signes aptes à servir de moyen de communication entre les individus ».

1.2 LES THEORIES DU LANGAGE : ENTRE LA PERCEPTIBILITE DU SIGNE ET L'INTELLIGIBILITE DU SENS

La présence du signe et du symbole en perpétuelle relation dynamique avec le sens a toujours été l'axe fondamental des polémiques entre tendances et écoles diverses. « Dans le rapport entre signe et sens, l'accent porte tour à tour sur le signe ou sur le sens, selon que le signe est le seul appui du sens, ou selon, au contraire, que la faculté d'appréhender par l'esprit quelque chose comme un sens explique que les signes fonctionnent comme signes »^[8].

[4] Jean Le Rond D'ALEMBERT. Discours préliminaire. In : Encyclopédie. Vol.1, 1751.

[5] F. De SAUSSURE. Cours de Linguistique Générale. Paris : Payot, 1915. Rééd. 1968.

[6] Otto JESPERSEN (1860-1943). Son grand traité *Langage* (1922) reste une grande œuvre d'analyses syntaxiques sur le prédicat, etc.

[7] Voir aussi son ouvrage *Précis de stylistique française*, 1946. 3^e éd., Paris, 1950.

[8] Paul RICOEUR. Signes et Sens. Encyclopædia Universalis, France S.A, 1995.

Le questionnement de la dualité signe/sens, bien qu'originnaire depuis l'apparition du stoïcisme^[9], n'a en réalité pris une forme empirique que depuis le XVII^e siècle après les périodes de l'antiquité et du moyen âge marquées par un discours philosophique reflétant la théorie platonicienne qui ramène les rapports entre signifiant et signifié aux simples relations idée et sens. En effet, le discours platonicien, tel qu'il se reflète à travers le *Cratyle*^[10], oppose deux approches sur la théorie des mots dont l'une affirme la naissance des mots par « convention » et l'autre avance la thèse qu'ils détiennent leur signification à partir du lien qu'ils ont avec la « nature ». Sa synthèse est « *qu'il faut aller aux choses mêmes « sans les mots* »^[11] partant du principe que la théorie du sens pour lui s'identifie à l'idée ou au « *principe intelligible aussi bien de la réalité que de la pensée* ».

Selon la théorie d'Aristote, le sens est en revanche porteur d'une notion de « forme » inhérente aux individus concrets. Cette approche qui s'oppose à la transcendance de la théorie platonicienne a de fait introduit la nouvelle dimension du *Concept* qui a su se maintenir jusqu'au Moyen Âge. « *Le concept n'est pas quelque chose que nous contemplons par l'esprit, mais que nous tirons par abstraction de l'expérience sensible* »^[12].

Il a fallu attendre l'an 1100, date d'apparition de la dialectique, deuxième des « arts libéraux »^[13], pour voir intervenir la grammaire dans la sphère linguistique. Pierre Abélard^[14] promulgue une nouvelle théorie globale du langage qui intègre la grammaire et la dialectique selon lesquelles les mots ne sont pas directement associés aux choses. Ils génèrent plutôt une « intellection » (*intellectus*) orientée vers ces choses, au langage d'agir en tiers pour reconnaître le réel. La grammaire est alors associée à la théorie de la signification que la dialectique peut lui conférer. Ainsi était envisagée l'apparition d'une science du signe que la grammaire spéculative^[15] a timidement développé en affirmant que le signe linguistique est exclusivement rattaché à la réalité mentale, qui, à son tour, conceptualise le réel. La grammaire devrait à cet égard partir d'une théorie structurale de l'activité psychique pour démontrer comment la langue reflète les structures de la pensée. Elle stipule que l'analyse des modes de signifié (*modi signandi*) est une véritable théorie des désignations basée sur le recensement des signes auxquelles l'analyse des modes de signification (*modi significandi*) ajoute un sens intentionnel qui met l'accent sur le rapport des « modes de désignation » et de « signification » aux « modes d'intellection » et aux

[9] La question du signe semble apparaître pour la première fois dans l'aire culturelle occidentale avec les stoïciens (III^e siècle av. J.-C.). Elle leur est nécessaire pour fonder la théorie logique très détaillée du syllogisme (la proposition conditionnelle ou disjonctive qui dévoile le mécanisme d'une preuve), articulée comme une série de signes sans signification propre, comme un système déductif opérant avec des termes (éléments initiaux) d'après des règles strictes.

[10] Ouvrage de Platon, du nom du philosophe héraclitéen (env.400-470).

[11] Paul RICOEUR. Signes et Sens. Op.Cit.

[12] « Le conceptualisme, qui est d'inspiration aristotélicienne, reconstruit l'ontologie en termes d'individus et de propriétés, interprétées intentionnellement. La contrepartie du concept dans la réalité (ce qui est visé par le concept dans le réel), c'est une propriété, non toutefois en tant qu'existant à l'état séparé, mais en tant qu'appartenant réellement à des individus. La relation exprimée par la proposition entre le prédicat et le sujet a pour fondement ontologique l'appartenance réelle de la propriété (visée par le concept que représente le prédicat) à un objet individuel (dénoté par le sujet) ». Source : Paul RICOEUR. Op.Cit.

[13] La base de toute approche de la connaissance était constituée par les deux groupes d'arts libéraux : celui des moyens de la pensée, de la compréhension et de l'expression (*trivium* : grammaire, rhétorique et dialectique), et celui des moyens de la connaissance du monde (*quadrivium* : arithmétique, géométrie, astronomie et musique, cette dernière étant avant tout l'étude de l'harmonie des choses).

[14] Né en 1079 à Pallet près de Nantes, France et mort en 1142. De tendance Platonicienne, il est reconnu pour son opposition à toutes formes de réalisme.

[15] École de grammairiens (env. 1250-env. 1350) qui a pu adjoindre la philosophie aristotélicienne à la théologie pour permettre de positionner le langage au sein d'une conception globale du savoir humain.

« modes d'être ». Durant toute l'époque médiévale, le courant nominaliste^[16] (*terminisme*) était la seule école ayant adhéré à un courant de pensée qui rapportent le sens au signe plutôt qu'à l'idée ou au concept.

L'apparition des mathématiques et la réorganisation de la méthode philosophique sur les principes mathématiques logiques ont engendré depuis le XVII^e siècle un recul dans le conceptualisme dominé par la physique aristotélicienne. On assistait depuis lors à un retours vers les idées mathématiques du platonisme et la philosophie cartésienne par le biais d'un renversement des principes des rapports du signe et du sens tels comme établis par le nominalisme et le conceptualisme^[17] du moyen âge. L'empirisme est ainsi parvenu par divers procédés et méthodes, diverses genèses et pratiques, à démontrer que les signes disposent d'un effet de substitution considérable; une substitution des choses mais aussi une substitution des signes les uns par les autres. Deux éléments données simultanément, l'un peut suggérer l'autre en sa présence ou en son absence voire même le substituer complètement. À notre sens, cette théorie associationniste de substitution aurait plus tard une part de concrétisation dans les méthodes et les systèmes de la codification et de la transmission électroniques des données reflétée à travers les théories de l'information et de la communication shannonienne. Selon les théories de Condillac^[18] « *les signes des langues sont d'institution, et non de nature* », de manière à ce que les relations qu'ils ont avec les idées qu'ils expriment soient totalement arbitraires et démunies de toute argumentation dans la nature même de l'idée ou de l'objet signifié.

1.3 LE PARADIGME DU « SIGNIFIANT/SIGNIFIÉ »

Les sciences linguistiques allaient ainsi établir avec Ferdinand de Saussure (1857-1913) et son ouvrage « *Cours de linguistique générale* » un débat profond sur la représentation sémantique et les rapports entre forme et contenu (ou signifiant et signifié). D'ailleurs, l'importance et les apports de Saussure dans le renouvellement des études linguistiques par le repositionnement de la linguistique proprement dite dans le vaste domaine de la sémiologie et la différenciation entre les rapports de synchronie et de diachronie, par la dimension opérationnelle concédée au concept de « système », par la théorie du signe et la théorie de la langue opposée à la parole ... constituent

[16] Jean LADRIERE identifie le nominalisme dans ces termes : « ... il ne reconnaît d'existence à aucune entité abstraite. Sous sa forme la plus extrême, il réduit les « universaux » au statut de simples noms et en fait donc des entités purement linguistiques. De façon positive, le nominalisme est un type de doctrine qui entend reconstruire entièrement l'ontologie en termes d'individus. Mais la notion d'individu peut évidemment être entendue de façon plus ou moins extensive. GOODMAN, qui a donné au nominalisme moderne l'une de ses formes les plus caractéristiques, présente le nominalisme comme une doctrine qui demande de traiter comme des individus toutes les entités dont l'existence est admise. La restriction ne porte donc pas sur le critère d'existence mais sur le mode de reconstruction de la réalité ». Source : Jean LADRIERE. Concept. In: Encyclopædia Universalis France S.A.1995.6-291c.

[17] Voir note 12.

[18] Etienne Bonnet DE CONDILLAC (1714-1780). Pierre TROTIGNON, professeur à l'Université de Lille III le présente dans ces termes : « *Examinant systématiquement toutes les formes de pensée, Condillac établit leurs liens respectifs avec le langage. Sans doute, on peut bien supposer une pensée sans signes, mais elle demeurerait enfermée dans des suites très restreintes d'idées ... Le langage est une invention purement humaine, non un don de Dieu ou de la nature. Les hommes, vivant en groupe, ont pu constater la liaison des actions avec des gestes ou des cris constants et ont établi un « langage d'action » qui a précédé le langage articulé. La substitution qui fit du langage articulé le moyen privilégié de l'expression et de la communication, à partir de l'état d'auxiliaire verbal du « langage d'action », est le résultat d'un long procès historique. Le grand mérite de Condillac est d'avoir exposé clairement des principes essentiels qui sont précisément ceux de la linguistique postsaussurienne : les signes des langues sont d'institution, et non de nature, en sorte que leur rapport avec les idées est arbitraire, c'est-à-dire sans justification dans la nature même de l'idée pensée ou de l'objet désigné* ». Source : Encyclopædia Universalis France S.A.1995.6-329 c.

le point de départ d'une révolution radicale, sur la voie de la linguistique structurale qui consoliderait la dissociation entre certains phénomènes perceptibles (les signifiants) pour désigner, évoquer, dénoter, signifier d'autres phénomènes non observables (les signifiés). La langue appartient en fait à la grande famille des systèmes de signes (l'écriture, l'alphabet des sourds-muets, les rites symboliques, la politesse, les signaux militaires, le système Morse, le Système Braille, la mode, les signaux maritimes, etc.). Le langage quant à lui est l'un des systèmes des signes sur lesquels aussi bien la tradition aristotélicienne que les définitions de la première moitié du XX^e siècle se basaient pour identifier les fonctions du langage.

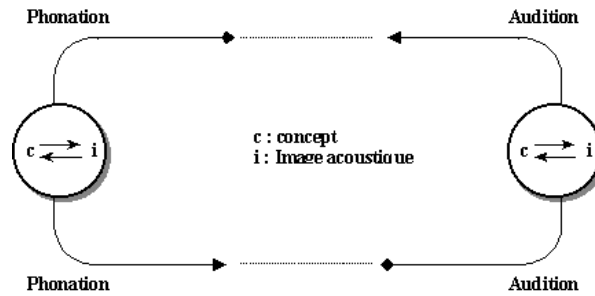


Figure 1 : Le schéma de Saussure sur la communication langagière

Bien que parti d'une autre approche sémiotique^[19], Saussure avait fini par admettre la version des stoïciens relative au signe verbal comme phénomène à double face constitué du *signifiant perceptible* et du *signifié intelligible*. « Un signe est un phénomène à double face qui oppose et relie un signifiant (vocal, écrit, gestuel, etc.) à un signifié corrélatif ». Il excluait de la sorte tout rapport à la chose qui va au-delà du domaine linguistique et ne considère qu'une différence interne au signe lui-même. Le signifié, dans ce sens, est purement et simplement une entité linguistique qui constitue la contrepartie du signifiant. L'interprétation de Saussure de cette doublure continuait à relever d'un ordre psychologique à travers lequel il qualifie le signifiant d'« *image acoustique* » et le signifié de concept ; tout deux puisés dans la langue elle-même qu'il considère comme un acquis « *inconscient déposé en chaque individu* ». Il n'hésitait pas cependant à considérer le rapport entre eux comme un mécanisme tout à fait arbitraire (le cas d'un signifiant et d'un signifié hétérogènes appartenant à deux ordres différents où l'aspect conceptuel du signifié n'est pas « motivé » par le caractère sonore, graphique ou gestuel du signifiant) malgré les remarques sur la « *motivation relative qui restreint l'arbitraire* », comme dans le cas de la composition et la dérivation des mots. En définitive, Saussure affirmait la linéarité du signifiant avec la même vigueur que l'arbitraire du signe ; une affirmation qu'on lui a reproché pour avoir confondu le rapport du signe à la chose avec le rapport du signifiant au signifié, partant du fait que ce dernier est porteur de plus de facteurs « iconiques » qu'il ne l'avait souligné.

La notion de l'information allait suite à ce débat voir se développer une tendance à distinguer entre le sens d'un message et sa forme. C'est en quelque sorte « une rupture mentale » entre deux aspects « étroitement complémentaires d'une même réalité » qui allait établir la distinction contenant/contenu pour permettre le profilage d'un secteur de recherche autonome qui se consacrera à la forme, en l'occurrence la sémiotique, chose à laquelle nous sommes redevables des réalisations actuelles en termes d'outils et de supports d'information et de communication électronique. Cette rupture allait s'accroître au début du XX^e siècle avec une distinction encore

[19] « Saussure partait d'une réflexion sur les régularités, voire les lois, découvertes par la linguistique comparée historique qui le conduisit à envisager le langage comme un système dont le signe serait la matrice constitutive ou l'élément fondamental. Partant de la langue en elle-même, Saussure précise la conception générale du signe à l'instar du signe linguistique ; ce dernier « résultant de l'association d'un signifiant et d'un signifié, nous pouvons dire plus simplement : le signe linguistique est arbitraire » ».

plus profonde traduite par une décomposition de la forme du message en deux sous classes : les symboles et les signaux qui constituent la forme réelle du support physique du message. De toute façon, le mot « information », bien que souvent utilisé pour faire référence au sens (contenu) d'un message, détient une origine étymologique qui le ramène toujours à l'idée de « forme ». Le mot latin « Informacio » veut dire « action de façonner », « donner une forme » et vient du terme latin forma,ae qui veut dire forme extérieure d'un objet. Le terme information dans une approche étymologique moderne, est également porteur d'une connotation de forme s'il est considéré d'un point de vue mathématique et technique.

1.4 SEMIOTIQUE ET SYMBOLISME DANS LA THEORIE LINGUISTIQUE

Avant d'aborder la théorie du signal dans le phénomène de communication, l'analyse du symbolisme et de la sémiotique linguistique requiert une considération préalable. La connotation iconique du langage allait en fait aboutir au développement d'une nouvelle théorie du symbolisme qu'Ernst Cassirer^[20] (1874-1945), puis Karl Bühler^[21] et Henri Delacroix allaient ensuite approfondir pour en faire tout un courant idéologique à partir du début du XX^e siècle. Bien que la nuance signe/symbole soit fondamentalement évoquée au niveau de l'arbitraire du signe et du conventionnalisme du symbole, cette théorie, également reprise dans une profondeur anthropologique par l'éthologie contemporaine, décrit la fonction symbolique comme pivot central du système symbolique de toutes les sociétés humaines. Claude Lévi-Strauss (1908-) le définit « *comme un ensemble de systèmes symboliques au premier rang desquels se placent le langage, les règles matrimoniales, les rapports économiques, l'art, la science, la religion* »^[22]. Un grand mérite revient dans ce sens à Charles Sanders Peirce (1839-1914) innovateur d'une science stoïcienne totale des signes qu'il appelait la *sémiotique* et qui devait couvrir tous les signes au-delà des signes linguistiques. Cette généralisation est intéressante étant donné que les lois du signe et du sens peuvent être appliquées à tous les autres ensembles sémiotiques (*rituels, règles de politesse, modes vestimentaires ou de l'habitat, symbolique de l'échange monétaire, classifications de tous ordres*) selon une variabilité fonctionnelle du signe qui peut prendre forme d'indice, d'icône ou de symbole. Cette dernière fonctionnalité symbolique du signe, retiendra notre intérêt par le fait qu'elle porte une dimension communicative stratégique dans les théories linguistiques et les systèmes d'information et de communication.

Emile Benveniste (1902-1976) le rappelle : « *Le langage représente la forme la plus haute d'une faculté qui est inhérente à la condition humaine, la faculté de symboliser. Entendons par là, très largement, la faculté de représenter le réel par un « signe » et de comprendre le « signe » comme représentant le réel, donc d'établir un rapport de « signification » entre quelque chose et quelque chose d'autre* »^[23]. Les contributions de Benveniste dans la consolidation de l'approche structurale de la linguistique par le ralliement de la notion de signe à celle de système signifiant allait révoquer en quelque sorte la théorie de Saussure selon laquelle il définit le signe comme association arbitraire d'un signifiant et d'un signifié. Selon Benveniste, du fait que le signe, comme d'ailleurs le réclame Saussure lui-même, est démuné de tout sens en dehors du système où il fonctionne, l'association entre signifiant et signifié doit être considérée à l'intérieur du système signifiant lui-même. Or, la constitution de cette association, une fois accomplie à l'intérieur du système, demeure interchangeable, donc en rien arbitraire.

[20] Voir ses ouvrages : La Philosophie des formes symboliques publié en 1923 (t. I, La Langue), 1925 (t. II, La Pensée mythique) et 1929 (t. III, Phénoménologie de la connaissance).

[21] Voir son ouvrage : *Sprachtheorie*. Vienne, 1934.

[22] C. Lévi STRAUSS. Introduction à l'œuvre de Marcel MAUSS. In : M. MAUSS, Sociologie et Anthropologie, éd. C. Lévi STRAUSS, préf. G. GURVITCH, 9e éd. Paris : PUF, 1985.

[23] Emile BENVENISTE. Problèmes de linguistique générale. 2 vol., Paris : Gallimard, 1966-1974.

La leçon de ces propos c'est finalement l'importance que revêt la médiation toujours présente dans tout phénomène de communication entre l'Homme et l'Homme et entre l'Homme et son entourage. Cet appareil médiateur, basé sur tout un système de symboles qui rend possible la communication de la pensée et la représentation de la substance de la réalité concrète, n'est autre qu'un ensemble de procédures (arbitraire ou conditionné), un mécanisme combinatoire de sens et de symboles, bref un « système » linguistique que les linguistes de tous les temps reconnaissent comme une extension naturelle des agrégats de l'expression de la pensée ou de l'acte de communiquer.

1.5 LA NOTION DE « SYSTEME » DANS LA PHILOSOPHIE LINGUISTIQUE

La notion de système en elle-même demeure très importante dans la considération du mécanisme de la langue et de la communication. Toujours en relation symbiotique avec la notion de signe et du symbole, la question du système comme mécanisme combinatoire d'éléments de désignation, de représentation ou de dénotation, est vue comme une science depuis la période des anciens stoïciens (III^e siècle av. J.-C.) et leur théorie du syllogisme^[24] fondée sur un mécanisme déductif d'une série de signes inductifs sans signification propre. Pour que cette logique fonctionne, il leur a été nécessaire de lui fournir une forme de « totalité réglementée » ... un système, une science où le signe et la sémiotique qu'ils engendrent constituent les bases d'une traduction systématisée.

La systématisation de la langue et de la théorie linguistique a connu son ampleur optimale à partir du début du XX^e siècle à la faveur des avancées empiriques, structurales et techniques de la langue. Le Cercle linguistique de Prague, une tendance fondée en 1926 par Vilém Mathesius (1882-1945), conjointement avec trois linguistes russes illustres S. Karcevskij, R. Jakobson et N. S. Troubetzkoy a marqué le démarrage véritable de la linguistique structurale du XX^e siècle à travers la publication de ses thèses dans les actes du I^{er} Congrès international de linguistique de La Haye en 1928. La conception de la langue comme système fonctionnel, primat de l'analyse synchronique, application à la diachronie^[25] de l'hypothèse de l'évolution convergente étaient parmi leurs points d'analyse fondamentale.

C'est ensuite au Cercle linguistique de Copenhague, avec Vigo Brøndal (1887-1942) et surtout Louis Trolle Hjelmslev (1899-1965), de soumettre la théorisation dans le domaine linguistique aux exigences de l'empirisme logique. Hjelmslev admet que le langage est un *procès* et que la sémiologie devrait lui faire rattacher un *système* capable de l'analyser et de le décrire par le biais d'un nombre limité de principes. Partant même du fait que la langue fait recours, tant au niveau du contenu que celui de l'expression, à tout un système logique combinatoire et à tout un réseau de relations de dépendance et d'interdépendance, il conçoit une algèbre de la langue, à l'écart de tout résidu psychologique ou sociologique.

[24] Défini par ARISTOTE, « Le syllogisme est un discours dans lequel, certaines choses étant posées, quelque chose d'autre que ces données en découle nécessairement par le seul fait de ces données » (Topiques, I, 1, 100 a 25 et Premiers Analytiques, I, 1, 24 b 18-20). Un syllogisme comporte donc un point de départ – les prémisses – et une conclusion. Celle-ci doit être nouvelle par rapport aux prémisses (ce qui assure la fécondité du syllogisme), mais elle doit, d'autre part, découler nécessairement de ces prémisses et d'elles seules (ce qui assure la rigueur de ce type de raisonnement). Source : Pierre OBENQUE. Syllogisme. Encyclopædia Universalis France S.A. 1995. 21-928 a.

[25] Les termes « synchronie » et « diachronie », introduits dans l'étude des langues par Ferdinand de Saussure, Par *synchronie* il entend l'état dans lequel se trouve une langue à un moment déterminé de son histoire. Il parle alors d'état de langue ou de *linguistique statique*. Par *diachronie* il entend l'évolution de la langue et parle alors de *linguistique historique* ou *dynamique*.

L'analyse de Saussure relative à cet aspect de système linguistique part plutôt d'une attitude envers les régularités et les lois mises en place par la linguistique comparée historique qui le conduit à définir le langage comme un système dont le signe serait « *la matrice constitutive* » ou l'élément essentiel, reprenant en cela la vision des stoïciens de la matérialité du langage. Les logiciens comme Rudolf Carnap (1891-1910) ont rajouté à cette vision saussurienne la définition du mécanisme linguistique en définissant la langue comme « *un système de signes avec les règles de leur emploi* ». Charles William Morris (1901 - ?) reprend la même définition dans son ouvrage *Signs, Language and Behaviour* (New York, 1946) où il précise que les langages doivent « *constituer un système de signes interconnectés, combinables de certaines manières et non d'autres manières* ». G. A. Miller (1956) en parle aussi comme un « *ensemble de symboles, et des règles pour leur utilisation* ».

1.6 LES PARADIGMES DE LA COMMUNICATION LANGAGIERE DANS LES THEORIES DE L'INFORMATION ET DE LA COMMUNICATION

L'apport le plus important à la définition de Saussure depuis les analyses de William Dwight Whitney^[26] (1827-1894), entièrement justifiées par la théorie de la transmission de l'information de Claude Elwood Shannon et Warren Weaver^[27], fut l'introduction du mot « communication » comme substitut à « l'expression de la pensée ». Une différence considérable dans l'analyse de chacune des deux méthodes est ainsi observée. L'expression de la pensée conduit inéluctablement à aborder les principes de l'introspection^[28], de la stylistique et de la logique alors que la communication engendre une « *observation scientifique du comportement communicateur et, plus encore, à l'observation du fonctionnement du système de communication, presque au sens technologique du mot « fonctionnement », comme c'est le cas dans la phonologie de Troubetzkoy^[29] et le fonctionnalisme de Martinet^[30] »^[31].*

Ce référent « technologique » du fonctionnement du procédé communicationnel allait en fait constituer le soubassement des principales théories de l'information et de la communication de ce siècle dont les plus admises sont celles de Shannon et Weaver.

1.6.1 La théorie mathématique de la communication

Shannon et Weaver, auteurs de *Théorie mathématique de la communication*^[32], œuvre de référence d'une envergure universelle sur la théorie de la communication, définiront cette même théorie sur les trois niveaux génériques d'un modèle bâti sur la trilogie émetteur, message et récepteur conformément au schéma détaillé suivant :

[26] Whitney marque nettement que le langage, né de et dans la communication dont il est l'instrument est fait de signes, que ces signes sont arbitraires et qu'ils s'organisent en structures pour former un système.

[27] Voir : C.E. SHANNON & W. WEAVER. *Théorie mathématique de la communication*. Paris : C.E.P.L., 1975.

[28] Observation d'une conscience individuelle par elle-même (Le Robert). Observation méthodique, par le sujet lui-même, de ses états de conscience et de sa vie intérieure, en psychologie (Larousse).

[29] Troubetzkoy ou Troubetzkoï Nikolai Sergueïevitch (1890-1938) s'est surtout consacré à l'étude des faits phoniques dans le langage. Héritier de Saussure et de Baudouin de Courtenay, il travaille à la fois sur l'hypothèse de la distinction langue-parole (Saussure) et sur celle de la distinction entre physiophonétique et psychophonétique (Courtenay), qui vont le mener à opposer la phonétique à la phonologie.

[30] André Martinet (1908-) peut être compté au nombre des principaux linguistes contemporains. Son œuvre, même en ce dernier quart du XX^e siècle où s'ouvrent à l'étude du langage des perspectives nouvelles par rapport à celles du structuralisme classique dont il est un des représentants, aura marqué son époque durablement.

[31] Georges MOUNIN. Linguistique : Objet et Méthodes. In : *Encyclopædia Universalis France* S.A.1995. 13-825 a.

[32] C.E. SHANNON & W. WEAVER. *Théorie mathématique de la communication*. Paris : C.E.P.L., 1975.

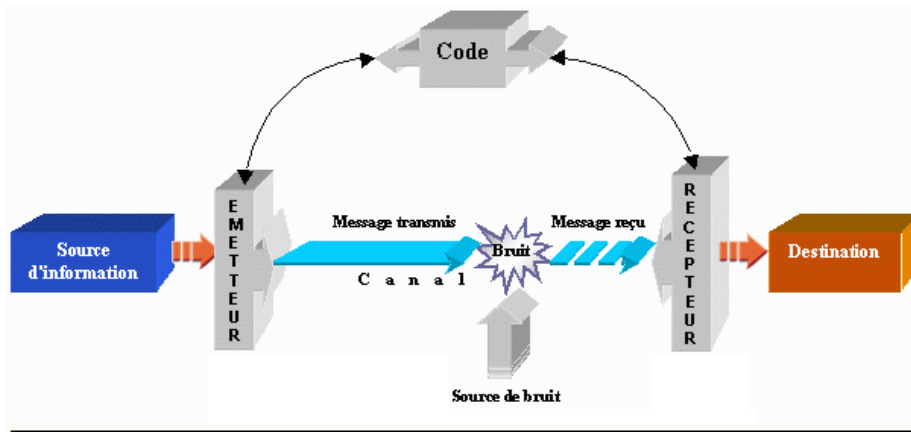


Figure 2 : Schéma de Shannon et Weaver, 1947

Selon ce modèle de communication, originaire d'un contexte de transmission de messages à distance par signal électrique, le processus de communication considère les opérateurs suivants :

- ❑ Une source d'information qui choisit un message particulier parmi tant d'autres messages possibles dans une forme orale, écrite ou graphique ;
- ❑ Un émetteur qui transforme le message en signal à transmettre sur un canal de communication. C'est un niveau où se pose le problème technique relatif à la précision avec laquelle les symboles de communication peuvent être transmis ;
- ❑ Un canal qui se trouve mis en forme par une perturbation engendrée par l'encodage du message réalisé par l'émetteur. La nouvelle forme du canal est donc celle des signes ou des signaux qui constituent les messages qui le parcourent. Un problème d'entropie^[33] est adressé à ce niveau pour évaluer la dégradation des signaux et la fidélité qu'atteindrait l'information transmise suite à son exposition à une source de bruit (interne ou externe) ;
- ❑ Un récepteur qui constitue en quelque sorte un émetteur dans le sens inverse dont le rôle est de changer le signal reçu en message compréhensible par le destinataire. C'est un niveau réservé à l'étude de l'efficacité avec laquelle le message reçu affecte-t-il la conduite du récepteur dans la compréhension du sens transmis.
- ❑ Une destination qui constitue l'aboutissement de la chaîne de communication. Le destinataire réagit face au sens du message conformément à deux éléments essentiels : d'une part l'état du message reçu suite à l'état de l'entropie et de la redondance qui en résultent, et d'autre part le référent culturel, c'est-à-dire la réalité non linguistique qui conditionne la compréhension du sens suite à l'élimination de la redondance dans le message.

Par cette stratification de l'acte de communication, chaque message peut alors être identifié comme une combinaison de signes porteurs d'une signification représentée elle-même dans un code particulier (arbitraire ou conventionnel). Ce qui implique en conséquence, que toute

[33] Introduite par Rudolf CLAUSIUS (1822-1888) il y a près de cent ans, reprise par BOLTZMAN et approfondie par Josiah Willard GIBBS (1839-1903), l'entropie « désigne en thermodynamique le degré de désordre d'un système physique (ce désordre selon le principe de Carnot, ne peut aller qu'en augmentant). En information, l'entropie devient le degré d'incertitude, mesurée en bits ou unités binaires, qui s'attache à une source émettant des messages. La néguentropie, soit l'inverse de l'entropie, décrit donc le passage d'un désordre aléatoire à un ordre relatif, et se trouve associée à l'organisation en général ». Source : Daniel BOUGNOUX. Sciences de l'information et de la communication. Paris : Larousse, 1993. p.407.

communication nécessite une source, différente ou non d'un émetteur (communicateur) qui code la signification fixée par la source dans un message, indifféremment de la nature du code utilisé qu'il soit des signaux acoustiques, vocaux ou non, signaux optiques, signaux directs non permanents (gestes) ou au contraire des signaux figés sous forme d'enregistrements ou de documents graphiques (écriture, photographie...). Il est important tout de même que les deux opérations de codage et de décodage s'effectuent dans un code qui leur soit commun. Le message ainsi mis en forme est acheminé via un support matériel, une voie de transmission ou un canal (air, ligne téléphonique, lettre, journal, etc.) vers un récepteur qui le déchiffre ou le décode afin d'en extraire, en fonction de son système référent, sa propre compréhension (version ou signification) du contenu sémantique. Le message circulant dans un canal est pourtant soumis à des perturbations aléatoires engendrées par le canal (endogènes) ou provenant d'une source externe (exogènes) à savoir bruit proprement dit, vent, distance élevée, inattention, distorsion du support en cas de documents papier, etc. qui engendrent une dégradation de l'information transmise. Ces altérations du signal, appelées bruit, sont compensées par le phénomène de la redondance qui « *consiste à répéter certains signes ou groupes de signes du message, ou encore, comme dans le texte écrit, à adopter un système où les mots contiennent davantage de signes que ne l'exigerait la communication de leur sens stricte* »^[34]. Les cybernéticiens, comme Norbert Wiener (1894-1964)^[35], ou les adeptes de la théorie générale des systèmes créée par Ludwig Van Bertalanffy^[36], ont intégré ce principe d'entropie dans le fonctionnement des machines auto-régulées en leur ajoutant un mécanisme de feed-back. En plus du canal habituel de l'acheminement de l'information entre les éléments d'entrée et de sortie, un canal parallèle appelé canal de rétroaction ou de feed-back, permet d'acheminer l'information de la sortie vers l'entrée dans une logique d'auto-régulation pour se maintenir et donc pour lutter contre la dégradation. Ce mécanisme de feed-back est largement utilisé dans les réseaux actuels de transmission des données : le contrôle de parité et le Cyclic Redundancy Check (Fig.8).

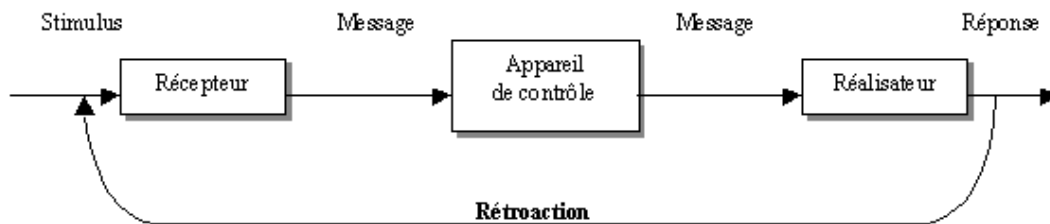


Figure 3 : Schéma de rétroaction de Bertalanffy

On identifie à travers cette structure du fonctionnement du mécanisme de communication le schéma élémentaire de la théorie de l'information qui constitue également un terrain d'analyse important pour la linguistique.

1.6.2 La théorie de l'information et la linguistique générale

L'idée fondamentale de la théorie de l'information, comme soubassement de *la théorie statistique de la communication*, est selon la conception shannonienne un croisement entre deux

[34] Thierry De SMEDT. Introduction aux technologies de l'audiovisuel : textes et schémas aux passages les plus complexes du cours de technologies des communications. Bruxelles : UCL Département de communication, 1996.

[35] Voir son ouvrage : *Cybernetics or Control and Communication in the Animal and the Machine*. New York : John WILEY & Sons, 1948.

[36] Biologiste théoricien, BERTALANFFY a attiré l'attention, à travers sa théorie générale des systèmes, sur la difficulté majeure que présente la reconstitution d'un tout intégré à partir des éléments constitutifs étudiés séparément.

courants en vue d'optimiser le transfert de l'information par l'intermédiaire des moyens de transmission (téléphone, télégraphe, télévision, etc.) via des canaux de communication (ligne téléphonique, ondes hertziennes...). Donc, d'emblée le mot information devient, selon Weaver « *une mesure de la liberté de choix dont on dispose lorsque l'on sélectionne un message [...] Le concept d'information ne s'applique pas aux messages individuels (comme le ferait le concept de signification), mais plutôt à la situation globale : l'unité d'information indique que l'on dispose, dans cette situation, d'une certaine liberté dans la sélection du message, qu'il convient de considérer comme un degré standard ou unitaire* »^[37].

D'une part, il s'agit d'une définition moderne de l'information proprement dite attribuée par les théories de traitement du signal et des dispositifs électriques de transmission pour étudier la quantité d'information, et analyser l'entropie d'une source d'information comme quantité mesurable par des formules mathématiques et logarithmiques strictes. D'autre part, il est question des propriétés des canaux qui servent à leur transmission (équivoque, transinformation, capacité, etc.) et la nature des rapports qui les unissent pour une optimisation de leur usage.

La théorie de l'information apparaît ainsi comme une théorie du signal au sens large qui intervient chaque fois qu'un signal est transmis ou reçu, et s'applique, donc à divers domaines comme la téléphonie, la télégraphie et le radar ainsi qu'à la physiologie du système nerveux ou à la linguistique. Daniel Bounoux affirme que « *pour se constituer mathématiquement, c'est à dire pour soumettre son objet à la mesure, cette théorie avait néanmoins pour préalable un sévère élagage de la notion courante d'information* »^[38]. Par contre, pour Michel Candelier [et.al.]^[39], c'est à Warren Weaver que revient en réalité le mérite d'avoir étendu l'articulation épistémologique du modèle Shannon-Weaver au delà du domaine purement technologique. Ils citent Weaver affirmant que « *Le mot « communication » sera utilisé ici dans un sens très large pour inclure tous les procédés par lesquels un esprit peut en affecter un autre. Ceci, naturellement, implique non seulement le discours écrit et oral, mais aussi la musique, les arts plastiques, le théâtre, la danse, et en fait tout comportement humain* »^[40].

Le point de rencontre de la théorie de l'information avec la linguistique serait défini à travers les deux phénomènes suivants : d'une part, le caractère probabiliste très liée à la statistique que donne Shannon à la quantité d'information véhiculée (calcul de fréquence d'apparition des unités du code dans un message, mesure de l'entropie...), et d'autre part la mesure de la redondance du signe linguistique. Toute information transmise est en fait soumise à un mécanisme d'appariement entre les éléments du message et ceux du code. L'économie générale de la communication consiste à minimiser les nombre de comparaisons possible qui dépend du nombre d'unités différentes du code ainsi que de la probabilité d'apparition de chaque unité dans le message. Cette réduction est généralement atteinte par l'établissement d'un ordre décroissant de fréquence d'apparition des unités dans le message. Le résultat de ce calcul pourrait servir à évaluer le taux d'information (proportionnellement inverse à sa probabilité) apportée par les unités dans le message en appliquant le principe de la redondance des unités linguistiques. Moins une unité est prévisible, plus elle apporte d'information. Donc, un signe devient redondant si le taux d'information qu'il rajoute est nul (probabilité maximale).

[37] Warren WEAVER. Contributions récentes à la théorie mathématique de la communication. In : Daniel Bounoux. Sciences de l'information et de la communication. Paris : Larousse, 1993. p.419.

[38] Claude SHANNON. Théorie mathématique de la communication. In : Daniel Bounoux. Sciences de l'information et de la communication. Paris : Larousse, 1993. p.407.

[39] Michel CANDELIER, Robert ESTIVAL, Yves GARNIER. Pour une archéologie schématique de la communication. In : Schéma et schématisation. N°19, 1983. ISSN : 0586-7606. Pp.9-20.

[40] Claude SHANNON, Warren WEAVER. The Mathematical theory of Communication. Urbana, 1949. (Citée par 38).

Le phénomène de la redondance des codes linguistiques, complémentaire à la théorie de l'entropie, revêt une grande importance à tous les niveaux du langage (sons, morphèmes, lexèmes, structures syntaxiques, écriture, etc.) ainsi qu'au niveau de l'économie du langage. Il s'agit de la décomposition du message selon des règles statistiques et linguistiques admises gérant l'usage des symboles constituant le message en question. « *On l'appelle redondance, car cette fraction du message est en fait redondante selon un sens proche de l'ordinaire; cette fraction du message et pas nécessaire (et par là se révèle répétitive) en ce sens que, si elle manquait, le message resterait complet ou du moins pourrait être complété* »^[41]. Dans des situations particulières où la redondance dépasse le bruit (beaucoup de perte d'information à valeur informationnelle nulle), elle résout des problèmes de sens par l'élimination du superflu. Un message bruité du genre « L. T.eor.e de l. .o.m.n.cat.on » est facilement restitué grâce à sa redondance élevée. Ce qui n'est pas le cas par contre pour une information statistique où chaque chiffre dispose d'une redondance souvent nulle (i.e. 1 23.6).

À ce phénomène est rajouté un principe fondamental des sciences linguistiques : le rôle joué par l'étude de la probabilité dans la constitution d'un message. Le choix des symboles successifs qui composent un message quelconque (oral ou écrit) du point de vue du système de communication, dépend entre autres du facteur de probabilité que les règles linguistiques d'une langue quelconque confèrent à son mécanisme morphosyntaxique.

En reprenant l'exemple de Weaver lui-même, « *en anglais, par exemple, il est évident que les probabilités exercent un certain contrôle sur le langage si l'on pense par exemple que le dictionnaire ne contient aucun mot avec la lettre initiale j suivie de b, c, d, f, g, j, k, l, q, r, t, v, w, x, ou z; de sorte que la probabilité est nulle pour que l'initiale j soit suivie d'une de ces lettres* »^[42]. Par contre, la probabilité pour que le mot suivant l'expression « Il est évident », soit le terme « que » est très grande mais non absolue, alors qu'après la chaîne de caractères « le moteur à combustion fonctionne par » la probabilité pour que le mot suivant soit « sable » est pratiquement nulle.

Ainsi, la théorie de l'information aurait permis de mettre au point une forme d'ouverture sur le domaine de la linguistique à travers le modèle général de la communication et de la théorie de l'information décrits auparavant. Même dans leurs aspects généralistes, ces modèles ont été utilisés par beaucoup de chercheurs pour décrire des faits linguistiques divers.

C'est d'abord le cas de Ferdinand de Saussure et sa théorie linguistique basée sur la description du circuit de la parole entre deux instances (deux individus, donc deux axes parallèles), qui communiquent par concepts et images acoustiques dans une symétrie de rapports entre les deux mécanismes de phonation et d'audition ([Fig.1](#)). Karl Bühler^[43] attribuait à son tour trois fonctions principales du langage rattaché au signe linguistique (l'énoncé) : une fonction d'expression du message (c.a.d. l'émetteur), une fonction de représentation (c.a.d. objets et états de fait), et une fonction d'appel qui constitue une action sur l'interlocuteur (récepteur).

[41] Warren WEAVER. Contributions récentes à la théorie mathématique de la communication. Op. Cit. p.421.

[42] Warren WEAVER. Contributions récentes à la théorie mathématique de la communication. Op. Cit. p.420.

[43] Voir son œuvre : Sprachtheorie. Vienne, 1934

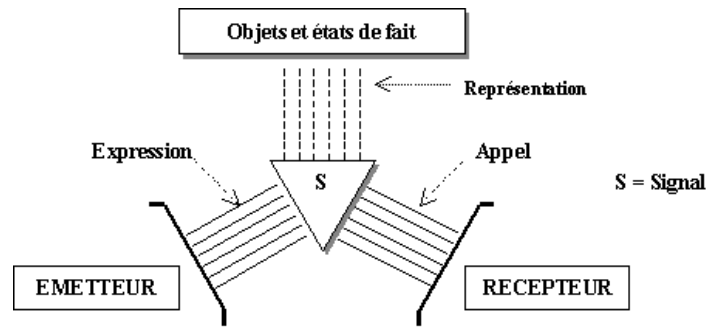


Figure 4 : Le schéma de Bühler

Selon Paul Ricoeur et Jean Sebestik d'une part^[44] et Pierre Kuentz de l'autre^[45], Roman Jakobson aurait mis au point un tableau de la situation de communication qui combine le schéma classique de Bühler et les schémas de la théorie de l'information structuré en six fonctions différentes du langage dont chacune s'identifierait à l'un des facteurs de l'acte de communication linguistique, en l'occurrence, l'émetteur, le récepteur, le canal de transmission, le code, le message et le référent. « Un destinataire envoie un message à un destinataire. Pour être opérant, le message requiert d'abord un contexte auquel il renvoie [...] ensuite [...] un code commun en tout ou au moins en partie au destinataire et au destinataire [...] enfin le message requiert un contact, un canal physique et une connexion psychologique entre le destinataire et le destinataire... »^[46].

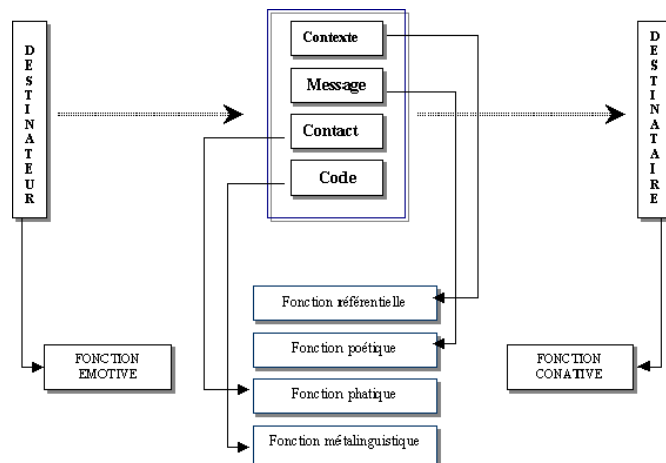


Figure 5 : Le schéma de communication et les fonctions du langage dans le modèle Jakobson

Ensuite Jakobson aurait établi une extension des six instances du processus de communication aux six fonctions linguistiques auxquelles elles donnent naissance. « Le message est spécifié dans sa nature par l'insistance que l'émetteur porte, soit sur ses propres affects (fonction émotive), soit sur les signifiés (fonction référentielle), soit encore sur le message particulier dans sa forme intrinsèque (fonction poétique), ou sur le destinataire (fonction conative). La fonction de prise et de maintien de contact avec l'interlocuteur (canal) et la fonction métalinguistique régulatrice du code sont des fonctions auxiliaires (Jakobson, 1963) »^[47].

[44] Paul RICOEUR, Jean SEBESTIK. Philosophie du langage. In : Encyclopædia Universalis France S.A.1995. 13-436 a.

[45] Pierre KUENTZ. Linguistique et littérature. In : Encyclopædia Universalis France S.A.1995. 13-857 b.

[46] Roman JAKOBSON. Essais la linguistique Générale. Paris : Minuit, 1963. Cité par Pierre KUENTZ. Linguistique et littérature. Op.Cit.

[47] Roman JAKOBSON. Essais la linguistique Générale. Paris : Minuit, 1963. Cité par Robert Page. Les processus de la communication. In : Encyclopædia Universalis France S.A.1995. 6-198 b.

1.7 DE L'ORALITE A L'ECRIT : ENTRE LA METAPHYSIQUE ET LA MATERIALITE DU SENS

S'il est certain que la linguistique constitue le modèle par excellence de la science générale des signes, tous les autres systèmes sémiotiques renvoient de façon ou d'une autre au langage, en tant qu'apothéose de l'esprit et de la pensée humaine. Tous les systèmes de communication et d'expression langagiers (phoniques, graphiques, iconiques, gestuels ...), préservant chacun ses propres caractéristiques, ne prennent leur autonomie propre et ne développent des caractéristiques particulières que sur la base d'une substitution de l'un à l'autre dans le même champ d'action de la langue.

L'écriture, l'une des formes encore les plus fortes d'usage dans la représentation des langues, malgré l'émergence explosive d'une nouvelle culture multimédiatique de la communication et de l'expression des idées, puise ses ressources, comme également affirmé par Saussure, dans sa substitution au langage parlé. En effet, selon Catherine Malabou, Saussure a démontré que « *parler revient à articuler entre elles des unités sonores et signifiantes en faisant jouer leurs différences* »^[48]. Cela signifie que cette différence de l'articulation — le cas des silences dans la parole ou la ponctuation dans l'écriture — constitue l'une des vérités intrinsèques du langage. À la lumière du modèle phonique, la notion de texte, fortement liée au concept d'écriture, considère le message écrit comme une articulation sémiotique : d'une part un signifiant matérialisé par des lettres dans un enchaînement en mots, en phrases, en paragraphes, en chapitres, en texte, et d'autre, part un signifié, « *sens à la fois originel, univoque et définitif, cristallisé par la correction des signes qui le véhiculent* »^[49]. La linguistique structurale aurait ainsi consacré scientifiquement la notion de signe dans son articulation signifiant/signifié et pourrait donc être appréciée comme un aboutissement réussi d'une métaphysique du sens.

Cependant, au delà de cette substitution acquise, l'écriture reste tributaire d'un système de signes spécifiques qui exige une structuration particulière du signifiant dans un mode qui ne pourrait être ramenée aux principes de l'oralité. En effet, l'aspect graphique donne aux signes, non seulement une substance (à l'instar de la substance phonique), mais également des caractères formels précis qui suscitent la garantie de la chose écrite et sa sauvegarde du sens originel de l'œuvre. Le texte est une arme de stabilité et de permanence de l'écrit contre l'usure du temps, de l'oubli, et contre les « roueries » de la parole, la fragilité et l'imprécision de la mémoire. L'inscription, en tant que telle, paraît comporter une distanciation spatiale irréductible à l'écart phonologique. Selon Jacques Derrida^[50] (1930-), l'écriture n'est plus fondée dans le langage parlé, mais, dans un sens plus stable et autant que possible unique. Le texte écrit gagne ainsi la définition d'une pratique signifiante, qui, appuyée par la sémiologie d'une part et par l'outil électronique de l'autre, concrétise de façon exemplaire la rencontre du sens et de la langue dans un univers d'information et de communication graphique et audiovisuelle complexe. Roland Barthes (1915-1980) semble également confirmer cette aliénation du texte écrit et lui confère une portée communicative importante en affirmant que « *nul ne peut prétendre réduire la communication à la simplicité du schéma classique postulé par la linguistique : émetteur, canal, récepteur, sauf à s'appuyer implicitement sur une métaphysique du sujet classique ou sur un empirisme dont la « naïveté » (parfois agressive) est tout aussi métaphysique* »^[51]

[48] Catherine MALABOU. J. DERRIDA. In : Encyclopædia Universalis France S.A.1995.

[49] Roland BARTHES. Théorie du texte. In : Encyclopædia Universalis. France S.A.1995. 22-372 a

[50] Voir son ouvrage : De la Grammatologie. Paris : Ed. de Minuit, 1965; L'écriture et la différence. Paris : Le seuil, 1967.

[51] Roland BARTHES. Théorie du Texte. Op.Cit.

1.8 ÉCRITURE ELECTRONIQUE : ENTRE LA DOUBLURE DE CODAGE ET LA VARIABILITE DE TRAITEMENT

Malgré les avantages du texte cités précédemment, le passage de l'oralité à l'écrit a constitué tout au début un détour important pour l'aboutissement à l'information. Le problème était d'exploiter un nouveau médium écrit dont l'efficacité et le sens reposeraient directement et essentiellement sur leur lecture. Selon Anne-Marie Christin et Jean-Pierre Balpe, bien que « *s'affichant sous une forme linéaire, où l'on a cru trop naïvement déceler une empreinte directe du discours, l'alphabet a perdu contact avec le sens* »^[52]. La projection auditive qu'il suggère, ne s'inspirent plus, comme c'est le cas de l'écriture idéographique comme « *révélateur verbal* », de l'aspect visuel des caractères qui le constituent. Se transformant en « *un véritable code phonétique, l'écriture alphabétique, il faut l'admettre, n'obéit pas, à la différence de celle qui se fonde sur l'idéogramme, à des principes de lisibilité* »^[53]. L'accès au sens par l'intermédiaire s'est en quelque sorte compliqué par la superposition d'une nouvelle couche entre le signifiant et le signifié ; une couche de transcription codée par des formes graphiques, en l'occurrence les lettres de l'alphabet, conçues et structurées selon un système de codage et de décodage des messages transcrits.

Ainsi donc, d'un stade où la représentation du sens était très rattachée à la langue à travers les représentations iconiques et l'écriture idéographique, l'écriture alphabétique s'est, quant à elle, éloignée davantage de la représentation phonique en adoptant un système de signes abstraits. Cette « refonte » de l'acte graphique cristallise une fois de plus le statut attribué à l'écriture en tant que système second basé sur l'adaptation des signes écrits à la langue. Sa forme graphique, tout en restant figurative (mais nullement iconique), est totalement indépendante de toute signification réaliste^[54]. Cette abstraction lui a valu une internationalisation et un plurilinguisme obtenus à travers la diffusion universelle de l'alphabet et de son usage généralisé pour quasi toutes les langues. « *Chaque langue s'efforcera désormais de distinguer sa propre prononciation des lettres, voire leur graphie, de celle des autres, se heurtant rapidement au fait que toute transcription de prononciation est condamnée à devenir caduque pour des motifs d'évolution interne ou encore, comme le montre le vietnamien romanisé, elle sera cause, si cette langue possède aussi une structure tonale, de complications graphiques invraisemblables* »^[55].

L'âge informatique allait consolider cette abstraction du caractère graphique, et même approfondir son écart du sens, par le report du principe de codage au-delà des limites d'une interactivité homme à homme et l'instauration d'une nouvelle couche de codification entre l'homme et la machine. Le codage de l'information électronique et la conception des interfaces Homme-Machine allaient suite à cette inférence technologique évoluer en parallèle comme soubassements fondamentaux des systèmes modernes d'information et de communication.

[52] Anne-Marie CHRISTIN & Jean Pierre BALPE. L'écriture. In : Encyclopædia Universalis France S.A.1995. 7-914
a

[53] Anne-Marie CHRISTIN & Jean Pierre BALPE. L'écriture. Op. Cit.

[54] C'est le cas des alphabets latin, arabe, chinois ... utilisés par des langues différentes sur le plan sémantique mais identiques sur le plan de la représentation graphique. Ces cas représentent la dissociation évidente entre le signe, le sens et le référent du texte écrit.

[55] Anne-Marie CHRISTIN, Jean-Pierre BALPE. L'Écriture. Op. Cit.

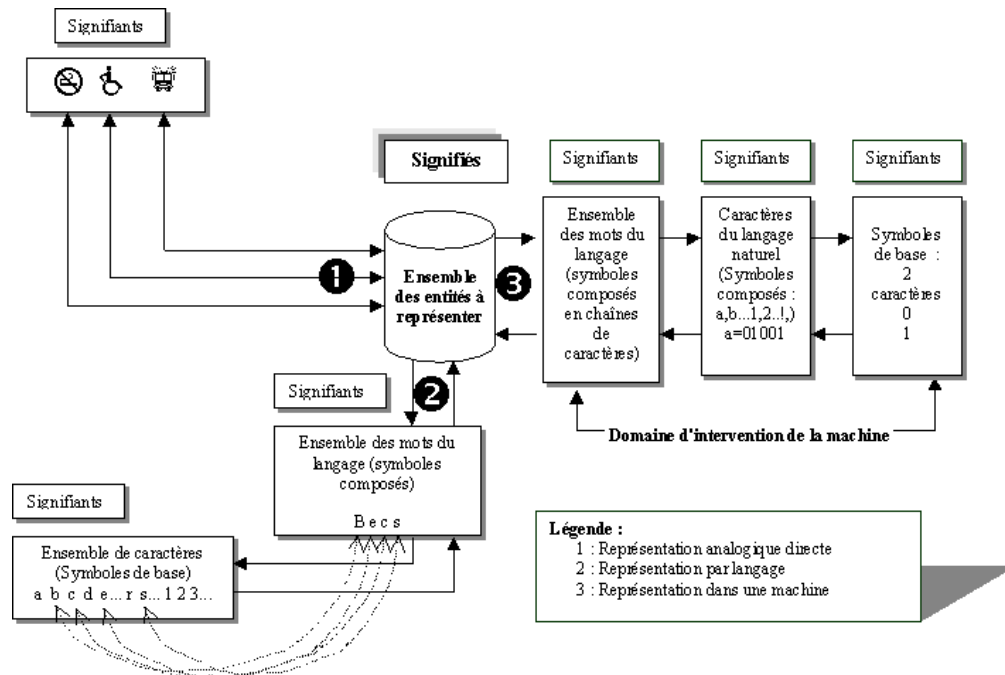


Figure 6: Synthèse des modèles Reix relatifs aux principes de la représentation de l'information

Les modifications engendrées par l'outil informatique sur l'écriture alphabétique sont en fait issues du domaine du traitement et de l'exploitation du signal électrique (voir [point 2 suivant](#)). Le principe de fonctionnement de l'électronique est basé sur le codage à deux temps (code binaire), représentés par les séquences de « 0 » ou de « 1 » tous deux équivalents à l'arrêt ou au passage du courant électrique. Partant de la pauvreté combinatoire de ces deux codes uniques et univoques, tous les signes traités par l'ordinateur (alphabétiques, phonique ou graphiques) sont réduits à des suites codées des 0 et des 1 accessibles au lecteur humain uniquement à travers un programme particulier de décodage. Les données codées acquièrent ainsi un aspect de virtualité ou d'immatérialité fortement tributaire d'un certain nombre d'opérations de calcul qui leur confèrent un aspect de perceptibilité aléatoire et fluide.

2. LA THEORIE DU SIGNAL

Le signal, dans sa définition générale est un terme que l'on retrouve dans la vie de tous les jours. « On parle de donner le signal de départ ou de lancer le signal de détresse, de signaler sa présence, etc. Ainsi l'idée du signal est associée à celle de la transmission d'une information dans le but de susciter une réaction »^[56]. Dans le domaine technologique, ce mécanisme de transmission par signal s'est largement basé sur le phénomène électrique comme source fondamentale de signalisation, toute abstraction faite de son contenu. Par signalisation, nous entendons dire la propagation des signaux sur un support de transmission quelconque. Dans le cas d'une signalisation électrique, la représentation de l'information à transmettre s'obtient par variation continue de la tension électrique (ou voltage), mesurée par rapport à un niveau de référence. Nous parlons de signalisation numérique lorsque les signaux à transmettre sont numériques et des signalisations analogiques dans le cas contraire.

[56] Samuel Pierre & Marc Couture. Télécommunications et transmission de données. Uébec : Télé-université, 1992. p.13.

2.1 L'ENERGIE ELECTRIQUE COMME SOURCE DE SIGNAL

Comme on s'en aperçoit tous les jours, le canal le plus utilisé dans la transmission, la diffusion et le traitement des signaux de communication est l'électricité malgré l'apparition ultérieure d'autres méthodes parfois plus efficaces comme la lumière (fibres optiques) ou le magnétisme (l'enregistrement magnétique, la radio) tous deux faisant recours partiel ou faisant référence à l'électricité comme source primaire d'énergie. La découverte de l'électricité et des procédés électromagnétiques avait cependant bouleversé le monde de l'information et de la communication par une variabilité innombrable de formes de support et de signaux de traitement des données.

Le traitement des signaux électriques a été en effet l'un des résultats considérables des opérations de recherches et d'observation de l'électricité menées depuis le XVII^e siècle sur des phénomènes de fluides électriques statiques et dynamiques. On peut sans doute attribuer l'origine de l'électronique aux débuts des années vingt du siècle précédent, quand Hans Christian Ørsted (1777-1851) a développé sa théorie de l'électromagnétisme observé par hasard à travers la désorientation d'une aiguille exposée au champ magnétique d'un fil conducteur parallèle.

C'est André-Marie AMPERE (1755-1836) qui met cette découverte en évidence sur une boucle parcourue par un courant électrique afin de transmettre une information à très courte distance. Une révolution importante a marqué depuis lors les modes et les méthodes de transport de messages bien que depuis avant cette date, le télégraphe de Claude CHAPPE (1763-1805) ou le télégraphe électrique de l'Américain Samuel MORSE (1791-1872) aient déjà révolutionné les *techniques de la communication à distance*.

James Clerk MAXWELL (1831-1879) résume quantitativement les lois de l'électromagnétisme à la fin des années 1860 et Alexander Graham BELL (1847-1922) invente la téléphonie sur fil en 1875 : il utilise les propriétés conductrices des fils métalliques pour transporter les variations de courant électrique produites par un microphone. Le XX^e siècle allait ensuite connaître l'émergence croissante du **travail par signes** et **sur signes** dans leurs natures les plus diverses (iconiques, fixes ou non ; linguistiques ; mathématiques ; etc.). L'électronique allait jouer un rôle prépondérant dans la floraison de tant de formes de technologies électromagnétiques toutes basées sur le nouveau principe du traitement du signal. La généralisation des méthodes de travail électroniques et électromécaniques des machines audiovisuelles, linguistiques, logiques et mathématiques allait consolider le principe de transmission de signaux sous forme de déclics et de déclenchement de processus selon la logique binaire du tout ou rien (marche/arrêt). Cette technique sera plus tard le soubassement fondamental des techniques de codage et du traitement du signal électrique.

La forme d'usage la plus connue de l'électricité comme support d'information a été sans doute la technique utilisée dans la transmission télégraphique en morse. Moyennant un circuit électrique comprenant une source d'énergie électrique, un interrupteur et une source lumineuse (substituée ensuite par une tête poinçonneuse), l'électricité provenant de la source d'énergie ne contient aucune information. Par contre, en actionnant l'interrupteur, il devient possible de donner au circuit électrique une certaine forme qui ne peut être que l'un des deux états : un courant qui passe ou un courant arrêté (les présages d'une technique binaire sur courant électrique!). Cependant, l'énergie électrique d'une source statique comme une pile ou des batteries fournissent des tensions dites continues qui ne changent pas de signe. Or comme il sera démontré à travers les découvertes ultérieures, pour être transmis ou manipulé, le signal électrique doit être émis sous une forme différente, plus adaptée à l'opération à effectuer : la transmission par radio ou l'enregistrement numérique par exemple qui constituent en tant que tels des opérations de codage

de l'information transmise par procédés analogiques ou numériques. Le courant alternatif constitue actuellement la source primaire d'une telle technique car issu de l'application d'un principe physique essentiel : l'induction^[57].

2.2 LE TRAITEMENT DU SIGNAL

Le traitement du signal occupe une place particulière dans la théorie de la communication. Il constitue le moyen par lequel la communication devient techniquement possible grâce au codage et au décodage des messages d'information sous forme de signaux qui relient un émetteur et un récepteur à travers un canal. C'est en quelque sorte le « *revêtement physique des messages [...] sous forme d'une grandeur de nature physique quelconque (acoustique, optique, électrique...) et généralement variable, par exemple au cours du temps* »^[58]. Cette variabilité est due entre autre au fait qu'un « *système de transmission n'est jamais en mesure d'émettre des signaux sans leur faire subir de déformations : selon leur nature, on parle de distorsion, d'affaiblissement, de diaphonie. Chaque type de support est caractérisé entre autres par son aptitude à transmettre un signal plus ou moins fidèlement* »^[59]. Un signal électrique doit toujours subir une transformation de forme conforme à une opération précise (i.e. transmission analogique ou enregistrement numérique) afin d'être transmis ou manipulé.

Dans l'univers de la communication actuelle, on parle le plus souvent de deux méthodes de codage et de décodage du signal électrique comme support d'information : le codage analogique et le codage numérique. Par codage analogique, on sous-entend une transformation continue du signal original et sa substitution au moment de la transmission par une autre onde électrique de fréquence élevée (i.e. émission radio). Par contre, un codage numérique consiste à découper un signal continu en petites unités discrètes appelées *Echantillons*. Le signal pourra alors être représenté par une suite de mesures successives sous forme d'une suite de nombres, semblables à un fichier de données informatique. On entrevoit dès maintenant combien les techniques de codage numérique marquent l'irruption définitive de l'information dans les technologies du son et de l'image.

Seulement, partant de l'abstraction qu'elle fait de la nature physique des grandeurs, la théorie du signal électrique est abordée conformément à plusieurs approches dont celles qu'Elie Roubine qualifie de signaux certains et aléatoires, à spectres continus ou à raies^[60].

- ♦ **Les signaux certains et les signaux aléatoires.** Selon la dite personne, « *la situation la plus courante est celle du signal certain, porteur de l'information, corrompu par un « bruit » de nature aléatoire. Toutefois, celui-ci peut être un parasite certain, la fréquence du secteur, ou un brouillage, par exemple. À l'inverse, l'information utile peut être portée par un signal aléatoire. C'est le cas de la radioastronomie, où l'on reçoit du « bruit dans du bruit »* ».
- ♦ **Les signaux à spectre continue ou à raies.** Partant de la propriété spectrale que détient le signal, en l'occurrence un support de sa transformé de Fourier^[61] (par référence à Joseph

[57] Le phénomène de l'induction peut être défini comme suit : « toute variation de tension électrique dans un conducteur engendre un champ magnétique (une aimantation) autour de ce conducteur, et toute modification du champ magnétique engendre une modification de la tension électrique présente dans un conducteur placé à l'intérieur de ce champ. » Source : Thierry De SMEDT. Op. Cit.

[58] Elie ROUBINE. Théorie du Signal. In : Encyclopædia Universalis France S.A.1995. 20-1071 b

[59] Jean Michel VANSTEENE. Modems-Howto. 13 février 1996. <http://www.guilde.asso.fr/miroirs/kheops/howto/Modems-HOWTO/Modems-HOWTO-1>

[60] Elie ROUBINE. Théorie du Signal. In : Encyclopædia Universalis France S.A.1995. 20-1071 c

[61] Procédé qui, par une technique d'échantillonnage continue ou périodique exercée sur le signal (mesures de différence de phases), permet d'identifier les différentes composantes qui le constituent.

Fourier (1768-1830)^[62], le spectre d'un signal analogique ou d'un signal discret (numérique) peut être respectivement identifié comme signal à spectre continu (i.e. lumière) ou signal à spectre de raies (périodiques). Dans un cas comme dans un autre, l'analyse spectrale peut s'accomplir en temps réels par l'intermédiaire de filtres^[63] sélectifs qui explorent la bande, visualisent le spectre sur un écran d'oscilloscope, prélèvent des échantillons par une méthode numérique pour mettre au point des calculs moyennant un ordinateur.

La théorie du signal converge actuellement vers l'analyse des modèles mathématiques grâce aux séries de Fourier^[64] qui représentent le signal électrique et renseignent sur ses modulations par les systèmes de transmission utilisés. Le signal est en fait indissociable des systèmes qui le transmettent et qui effectuent sur lui l'une des deux fonctions majeures de traitement (amplificateurs, modulateurs, détecteurs, etc.) ou de détérioration (bande passante insuffisante, bruit). À l'instar de tout processus de communication où toute mesure effectuée est altérée par des erreurs inhérentes à l'émetteur et au récepteur, dans tout système de télécommunication, le signal reçu est également détérioré par le canal qui l'a transmis. D'où la nécessité de recourir à un traitement du signal pour remettre le message à son état initial. Les techniques de traitement convergent désormais vers la solution d'une amélioration en temps réel du signal. C'est-à-dire que l'analyse spectrale maintenant possible en temps réel, couvre une panoplie d'applications allant de la « *caractérisation d'un signal par sa composition et sa densité spectrale jusqu'à l'auscultation des structures, moteurs, détection de défauts par comparaison à un matériel vibrant jugé correct* »^[65]. Partant de ce principe de dynamique et d'instantanéité dans le traitement du signal, on se limite généralement à une méthode de traitement linéaire entre un signal intrant (reçu) et un signal extrant (traité). Si auparavant l'analyse spectrale s'effectuait instantanément à travers des filtres passe-bande^[66], sorte de dispositifs électroniques qui restreignent l'espace de fréquence attribué à la transmission du signal sur une liaison donnée, la théorie de la *réalisation*, issue elle-même de la mise en place physique des filtres, a donné lieu à des méthodes de contrôle effectuées par des ordinateurs capables de réaliser un filtrage échantillonné par algorithmes spécialisés. Ces méthodes ont démontré que pour représenter un signal sans ambiguïté et le reconstituer fidèlement (donc, conserver le maximum d'informations qu'il véhicule), il est nécessaire de procéder à un échantillonnage du signal à une fréquence double de celle la plus haute qui caractérise le signal. En définitive, un signal analogique (i.e. en téléphonie) est soumis à trois opérations distinctes : échantillonnage, quantification^[67] et codage; toutes trois relevant de la théorie de l'information.

3. LA REPRESENTATION DES DONNEES ET LE CODAGE DE L'INFORMATION

Tout phénomène de communication n'est donc rendu possible que s'il s'effectue sur la base d'un code commun non seulement porteur d'un sens, mais en plus compréhensible pour l'émetteur et

[62] Joseph FOURIER, mathématicien et physicien français qui aurait amélioré les techniques d'approximation des racines des équations polynomiales. Son œuvre originale est sa théorie sur la propagation de la chaleur dans un solide.

[63] On applique au signal de départ un *filtre passe-bande* qui restreint l'espace de fréquence attribué à la transmission du signal sur une liaison donnée.

[64] Méthode d'analyse qui ne s'applique qu'aux cas particuliers des fonctions périodiques.

[65] Georges NEY. Mesures électriques et électroniques. In : Encyclopædia Universalis France S.A.1995. 15-41 c

[66] Les filtres passe-bande atténuent les fréquences situées à l'extérieur d'une gamme de fréquence appelée bande passante du filtre et transmettent sans atténuation notable les fréquences comprises dans cette bande. Source : Samuel Pierre & Marc Couture. Télécommunications et transmission de données. Québec : Télé-université, 1992.p.36.

[67] La quantification, selon André ROUMANET et Jean-Luc VERLEY (cf. Théorie des ensembles. In : Encyclopædia Universalis France S.A.1995. 8-421 c), est une méthode qui associe au signal analogique un nombre d'états possibles généralement élevé (212) alors que le traitement numérique associe à chaque état de quantification une suite de valeurs booléennes estimée à 12 variables.

le récepteur à la fois. On parle alors d'une forme particulière de message appelé *codage* ou *modulation* issue de l'application d'un *protocole*.

À travers l'acte de communication humaine, nous utilisons des éléments linguistiques qui constituent dans une stratification verticale un alphabet, un vocabulaire, une syntaxe et une sémantique... bref une langue. En fait la langue n'est qu'un aspect d'un ensemble prédéfini de codes ou de signaux complexes certes, mais chargé d'un sens pour tous ceux qui l'adoptent. Ces faits ont été souvent utilisés par les linguistes et les cybernéticiens pour mettre en évidence la ressemblance entre tout acte de communication (même si émetteur et/ou récepteur sont des machines) et le langage.

Coder une information c'est en réalité lui attribuer une valeur représentative autre que sa forme d'origine compréhensible par les procédés qui ont contribué à sa genèse. Depuis l'aube des temps, les graffitis et les peintures rupestres, les hiéroglyphes, les signaux de fumées des amérindiens, les tam-tams africains ou même les techniques du code Morse... se basaient sur ce principe de représentation progressivement abstraite pour coder un sens, une information transmises à distance et compréhensible grâce à un référent commun de part et d'autre par l'émetteur et le récepteur.

Ceci sous-entend qu'il existe une mémoire partagée, un référentiel commun dans lequel sont maintenues toutes les formes possibles utilisées pour réaliser l'opération de l'encodage, également appelée *production*. Les conditions dans lesquelles ce mécanisme s'accomplit (engrammes innés, apprentissage fondé sur des lois statistiques de probabilité...) ne sont pas encore définitivement tracées pour être dépendantes de beaucoup de facteurs culturels et affectifs aléatoires. Bref, toutes les instructions (phonation, mimiques et comportement gestuel) sont transmises à l'appareil moteur à partir d'un code. Et c'est dans un sens inverse mais non symétrique que le récepteur effectue OBLIGATOIREMENT l'opération de décodage, également appelé *reconnaissance*. L'obligation émane du fait qu'il est inconcevable qu'un message soit uniquement axé sur le code. D'ailleurs, il faut se rappeler que chacun des deux concepts message/code n'a d'existence que par rapport à l'autre. Le code ne peut donc être appréhendé autrement qu'un ensemble de conventions auxquelles émetteur et récepteur d'une communication par message font recours pour les deux opérations de production et de reconnaissance.

Devant l'aspect aléatoire et restrictif dans le recours à des conventions de codage et de décodage entre émetteur et récepteur de messages, certains usages ont trouvé ces procédés assez démunis d'un minimum de cohérence et de sécurité de transport. Les contextes militaires, diplomatiques ou même économiques ont contribué à mettre en place des techniques variées pour simplifier le codage de l'information sur une base beaucoup moins arbitraire et plus sécurisante. Le codage binaire allait donner une nouvelle dimension aux techniques de la représentation de l'information et les « programmes » de son traitement et de son analyse.

Ramenés tout particulièrement au domaine de la transcription graphique, en l'occurrence la communication écrite qui constitue notre domaine d'intérêt actuel, les mécanismes de codage et de décodage des messages écrits passent sous la houlette de cette conception de représentation initialement appréhendée sous forme scripto-graphique, puis alphabétique pour subir finalement les métamorphoses du support électronique et l'adjonction d'une nouvelle couche de représentation du sens : la couche électronique (analogique ou numérique).

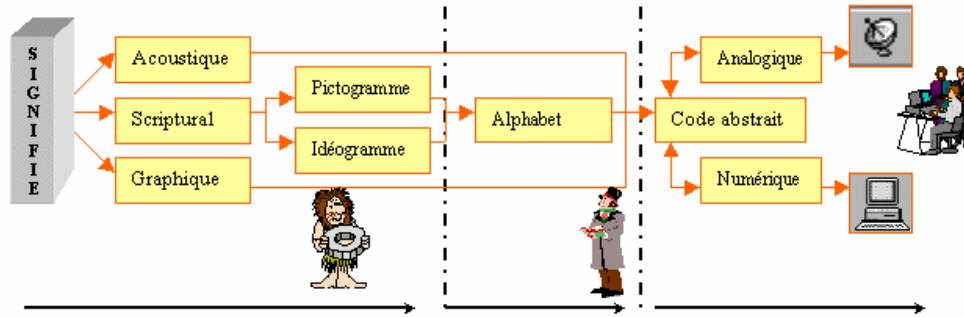


Figure 7 : L'évolution des couches interfaces information/homme

Il s'agit d'une dématérialisation plus profonde et d'une virtualité de l'écrit électronique qui, selon Christin et Balpe, transmettent au texte des caractéristiques que l'écriture ne détient au niveau d'aucun autre support. En effet, la codification binaire des signes étant soumise à un système unique de codage, la transition souple et facile entre un code et un autre permet une variabilité d'affichage engendrée par les modifications des procédures de calcul de leurs codages. Beaucoup de systèmes proposent actuellement la possibilité de changer les polices de caractères utilisées, les fontes et la direction de l'écriture (i.e. le cas d'un multilinguisme lourd^[68]).

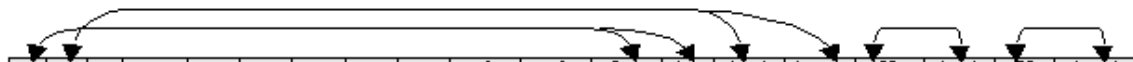
Dans le contexte informatique, l'unicité des codes binaires, unités insignifiantes mais non figées, permet, grâce à une série de lois externes, une générativité et un traitement intelligent du texte par le biais de produits devenus désormais courants comme les correcteurs orthographiques ou syntaxiques.

3.1 LE CODAGE BINAIRE DANS LES SYSTEMES DE TRAITEMENT ET D'ECHANGE DES DONNEES

Le codage binaire constitue en fait le soubassement qualitatif fondamental de la théorie de l'information de Shannon ; une théorie qui se basait sur la quantification de l'information codée en bits donc soumise à des règles de quantification mathématiques strictes inspirées de l'application de la logique Booléenne plutôt que de la simple quantification du transport électrique des signes et symboles. La logique du binaire allait en effet trouver son plein essor grâce à la logique de l'algèbre de George Boole (1815-1864) mise au point en 1847 et servant de prototype pour plusieurs domaines comme la théorie des ensembles, la logique algébrique, la notion de spectre d'anneau, la théorie de la mesure et du calcul des probabilités, dans les domaines de l'informatique, de l'indexation et de la recherche de l'information (ET, OU, SAUF). À son origine, et conformément à la théorie des ensembles, l'algèbre de Boole applique sur un ensemble « A » non vide, deux lois de composition interne « V » et « ∧ » (associatives et commutatives) et une application unaire « ¬ » moyennant deux éléments prioritaires le « 0 » et le « 1 » pour vérifier les axiomes suivants :

1. $\forall x \in A, \forall y \in A, (x \wedge y) \vee y = y$;
2. $\forall x \in A, \forall y \in A, (x \vee y) \wedge y = y$;
3. $\forall x \in A, \forall y \in A, \forall z \in A, x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$;
4. $\forall x \in A, \forall y \in A, \forall z \in A, x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$;
5. $\forall x \in A, x \wedge 1 = x$;
6. $\forall x \in A, x \vee 0 = x$;
7. $\forall x \in A, 0 \vee x = x$;
8. $\forall x \in A, 1 \wedge x = x$;

[68] J'oppose multilinguisme « lourd » (codage, typographie, morphologie, syntaxe, sémantique et surtout direction d'écriture), au multilinguisme « souple » (sans ce dernier critère). Voir <http://www.cnet.fr/ust/bhenda.html>



x	y	z	x+y	x.y	y+z	y.z	x+z	x.z	x.^x	x+^x	0+x	1.x	(x.y) + y	(x+y) y	X (y+z)	(x.y) + (x.z)	X + (y.z)	(x+y) + (x+z)
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
0	1	0	1	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0
0	1	1	1	0	1	1	1	0	0	1	0	0	1	1	0	0	1	1
1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	1
1	0	1	1	0	1	0	1	1	0	1	1	1	0	0	1	1	1	1
1	1	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

Tableau 3: Tableau de validation des axiomes mathématiques de Boole par des éléments binaires.

La théorie de Shannon allait également trouver son inspiration dans la thermodynamique pour aborder l'analyse de l'entropie (dégradation d'un signal en présence d'une source de bruitage) et par conséquent proposer des modalités à son contournement (Fig.8).

	0	1	1	0
1	1	0	0	0
1	0	1	1	1
1	1	0	1	1
1	0	0	1	0

Figure 8 : Modèle d'analyse de l'entropie de Shannon

Source : Philippe Breton : Une histoire de l'Informatique. Édition La Découverte, 1990, p.49.

Commentaire : L'entropie, étant définie comme une dégradation possible d'un signal au moment de la transmission suite à une infiltration d'une erreur de codage issue d'un effet de parasite, le mécanisme des chiffres de contrôle assignés à chacune des lignes et des colonnes de la matrice de codage, permettra de vérifier l'exactitude de la reconstitution du message une fois les données arrivées à destination. Ce mécanisme connu actuellement dans l'univers des télécommunications sous l'appellation de *contrôle de parité*. Il s'agit d'un processus simple utilisée dans les opérations de transfert des données et de la communication à distance qui sert à reconnaître les erreurs de transmission ou de stockage de données. Il consiste à ajouter un bit de parité pour chaque octet et à le vérifier. S'il y a une erreur, le bit de parité transmis sera différent de celui que l'on a calculé. Par ce processus, il n'est pourtant pas toujours sûr qu'il n'y a qu'une seule erreur par caractère, car deux fautes dans le même code peuvent ajuster le calcul tout en biaisant le contenu.

Dans l'exemple proposé, il est évident que le chiffre parasite est un 0 car avec une valeur 1 le total de la ligne et de la colonne correspondantes seraient de la valeur paire, soit 0.

Cette technique est progressivement remplacée dans ses procédures de fonctionnement et non dans sa structure opératoire (toujours un contrôle de parité), par des techniques beaucoup plus sophistiquées et dynamiques. La technique « CRC » abréviation de « Cyclic Redundancy Check = contrôle de redondance cyclique » consiste à effectuer des contrôles itératifs au moment de la transmission des données ou de leur enregistrement sur support magnétique pour déterminer s'il y a une erreur. Si tel est le cas, une retransmission instantanée est effectuée pour rectifier l'erreur insérée accidentellement.

L'inventeur du code binaire est l'Anglais Sir Francis BACON (1561-1626) dont le souci de crypter des messages diplomatiques l'a poussé à inventer un système de transcription qui assigne à chaque caractère alphabétique une représentation composée de deux symboles (alphabétiques dans ce cas). À chaque symbole, une représentation typographique supplémentaire est attribuée pour garantir un maximum de complexités de déchiffrement des textes transmis (Fig.9).

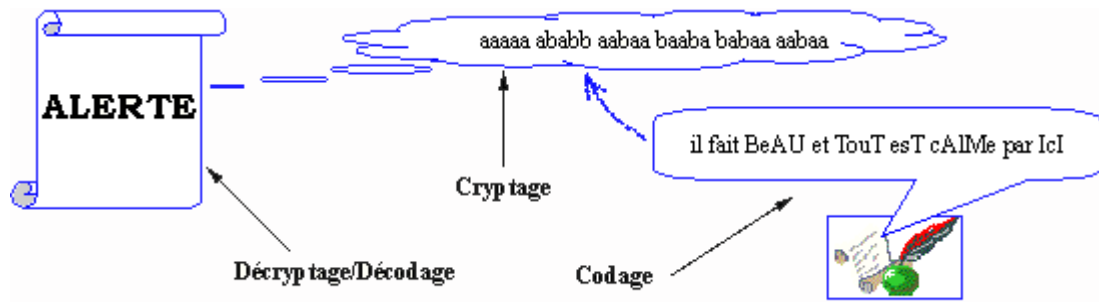


Figure 9 : Exemple de codage en binaire alphabétique selon le modèle BACON

Commentaire : Selon le mécanisme de F.BACON, à chaque caractère est attribuée une représentation binaire composée de 2 symboles graphiques (a et b). Dans un codage pondéré à 5 positions binaires, la lettre « A » serait alors équivalente à « aaaaa », la « B » équivalente à « aaaab », la « C » à « aaaba »... et la « Z » serait équivalente à « bbaba ». L'effet de cryptage serait dans un deuxième niveau de codage apporté par une typographie différente assignée à chaque symbole « a » ou « b » de sorte que les minuscules seraient codées en « a » et les capitales seraient codées en « b ». Cette technique convertirait le message « il fait BeAU et TouT esT cAlMe par Ici » en codes binaires (alphabétiques) « [aa aaa][a babb] [aa baa][b aab a][baba a][aa b aa] » qui à son tour est traduit en « ALERTE ».

Il a fallu attendre les années 1800 pour que l'une des premières combinaisons de ce procédé binaire avec la plate-forme automatique prenne forme d'un métier à tisser que Joseph-Marie JACQUARD (1752-1834), mécanicien français, a inventé, perfectionné et achevé en 1806. « *Le moulin à chiffre* » de Charles BABBAGE (1792-1871), qui pourrait être considérée comme l'ancêtre des ordinateurs contemporains ou « *la machine de Turing* » d'Alan Mathison TURING (1912-1954) conçue en 1936 et considérée actuellement à la base de toute la théorie des automates et, d'une manière générale, de la théorie de la « calculabilité »... constituaient des formes de continuité à ce couplage stratégique entre les procédés binaires et l'automatisme.

L'histoire des techniques surtout celle versée vers la logique des calculs, a été, comme la trace si bien Philippe BRETON^[69] dans son ouvrage « *Une Histoire de l'Informatique* », parsemée d'inventions de « *machines à calculer* » de tout genre : depuis les calculateurs analogiques de Vannevar BUSH (1890-1974) aux calculateurs numériques électromécaniques de George STIBITZ, de Howard AIKEN (1900-1973) et de Konrad ZUSE (qui eu le mérite en 1944 de mettre en place le premier ordinateur numérique opérationnel basée sur la technique binaire numérique contrôlée par un *programme affiché* adaptable selon les opérations à effectuer) en passant par les calculateurs numériques électroniques de John V. ATANASOFF, de J.Presper ECKERT et John W. MAUCHLY.

La grande révolution dans le domaine des calculateurs électroniques qui a marqué jusqu'à nos jours le domaine désormais appelé « *l'univers informatique* » a eu lieu avec la nouvelle structure de machines à *programmes enregistrés*, par opposition aux machines des générations antérieures basées sur des *programmes affichés*, dotées d'une mémoire capable de contenir à la fois les programmes et les données à traiter.

En effet, il a été conclu que la phase d'un automatisme complet ne s'accomplirait que si les programmes qui gouvernent l'exécution des calculs pouvaient être enregistrés avec les données à l'intérieur de la même machine. Ceci a engendré le recours au principe d'algorithmique, c'est à

[69] Philippe BRETON. Une histoire de l'informatique. La Découverte, 1990.

dire l'enregistrement simultané dans la mémoire de la machine d'une série de tâches à réaliser et des données elles-mêmes. C'était le début de l'usage des programmes enregistrés en codes binaires numériques^[70].

3.2 LE CODAGE BINAIRE : LES AVATARS DU NUMERIQUE

La structure de ces machines, dont le premier modèle était présenté par le projet EDVAC^[71] (Electronic Discrete Variable Computer = Ordinateur électronique à variables discontinues) a été décrite en 1946 par un mathématicien américain d'origine allemande, John Von NEUMANN. Tous les systèmes informatiques actuels portent d'ailleurs son nom : « *Architecture de Von NEUMANN* » qui formalise la distinction entre le hardware et le software et qu'il serait important de détailler par souci d'exposer le nouveau mécanisme de codage et de traitement des données que nous aborderons dans une grande partie de ce travail.

C'était en 1954 que John Von NEUMANN et ses collaborateurs ont défini l'architecture de base que l'on utilise encore aujourd'hui dans la plupart des ordinateurs. C'est une architecture basée sur les principes suivants :

- ❑ Les ordinateurs sont structurés sur une base modulaire en unités fonctionnellement indépendantes : une unité de commande, une unité de calcul (destinée à exécuter les opérations arithmétiques et logiques du programme en cours d'exécution), une mémoire interne et des unités périphériques ;
- ❑ Un espace mémoire central (RAM), dont le rôle est de contenir temporairement et cote à cote les programmes de traitement ainsi que les données traitées. Cette mémoire interne est divisée en cellules auxquelles on peut attribuer des adresses (en entiers naturels séquentiels) ;
- ❑ Chaque cellule contient une unité d'information et une seule ;
- ❑ Le système binaire est utilisé pour représenter les commandes et les données ;
- ❑ Les commandes et les données sont stockées dans la même mémoire interne ;
- ❑ Les cellules dans lesquelles sont stockées les commandes successives ont des adresses qui se suivent ;
- ❑ Les commandes sont exécutées en séquences ;
- ❑ La séquence de traitement peut présenter des interruptions conditionnelles ou inconditionnelles.

[70] L'adjectif « *numérique* » signifie à peu près « *dénombrable* ». C'est le contraire d'analogique. Tandis que les quantités analogiques peuvent prendre toutes les valeurs intermédiaires et sont donc divisibles à l'infini, les valeurs numériques ne peuvent avoir qu'un nombre limité, discret, de valeurs intermédiaires. En anglais, numérique se dit « *digital* » et l'étymologie latine est « *digitus* » (doigt) car, depuis toujours, nous utilisons nos doigts pour compter. On appelle numérisation la transformation d'une valeur analogique en une valeur numérique ; l'appareil ou le composant du système qui effectue cette transformation est le transducteur analogique-numérique. L'adjectif « *analogique* » signifie « *qui se suit et se ressemble* ». Un signal analogique est une quantité physique qui, à l'intérieur de limites données, peut adopter n'importe quelle valeur, c'est-à-dire un nombre infini de valeurs intermédiaires. Par exemple, les aiguilles d'une montre à affichage analogique indiquent l'heure en passant par tous les angles possibles entre 0 et 360 ; à chaque instant leur position est différente. Une montre à affichage numérique n'a pas d'aiguille et n'indique que des valeurs intermédiaires discrètes, celles des différentes heures et minutes.

[71] L'EDVAC I était équipé d'une mémoire pouvant contenir 1000 mots. Il pouvait effectuer les opérations arithmétiques de base, mais ne pouvait pas, en revanche, effectuer de contrôle d'erreurs. EDVAC II pouvait exécuter les opérations arithmétiques de base à une décimale avec virgule fixe et effectuer automatiquement le contrôle d'erreurs. EDVAC III incluait en outre le calcul automatique à virgule flottante. Sa mémoire passait à 4000 mots.

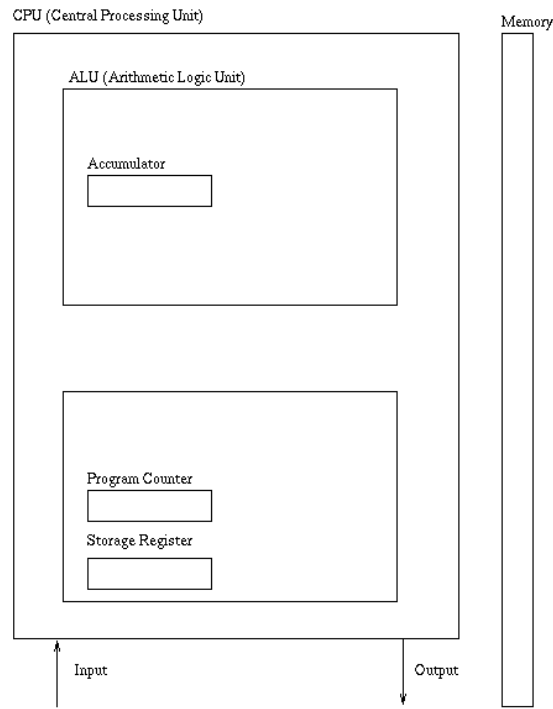


Figure 10 : L'architecture informatique de Von Neumann

Les informations traitées sur une base binaire numérique sont représentées par des grandeurs qui ne peuvent être que de deux états distincts qui s'annulent mutuellement. À chacun de ces états est associée, ***par convention***, une valeur unique, 0 et 1 (binaire numérique = Binary digIT = BIT) de l'élément correspondant. Dans tous les systèmes utilisant le système binaire, les informations sont représentées à l'aide d'une codification à deux états uniques. Dans les systèmes binaires numériques, l'information élémentaire est constituée par un ensemble de bits dont la position relative joue un rôle important dans la cohérence du système de représentation (codage) approprié (à [pondération](#) ou [réfléchi](#)).

Par une transposition de ces grandeurs de représentation sur le contexte de l'information, l'émetteur a la possibilité de coder son message de telle sorte que les signes Zéro (0) et un (1), équivalents à une ouverture ou à une fermeture d'un circuit (en l'occurrence électrique), constituent les deux alternatives conventionnelles strictes. La quantité d'information est donc évaluée par le logarithme^[72] du nombre de choix possibles au moment de la représentation binaire. Dans ce cas, l'information, codée sur le principe binaire est proportionnelle à la formule suivante :

$$\text{Quantité d'information} = I = k \lg(N/n)$$

où N est le nombre d'événements possibles et n le sous-ensemble désigné par l'information. Pour attribuer à la quantité d'information les propriétés des grandeurs mesurables, on considère par définition l'élément k comme constante qui dépend du choix de l'unité. En prenant comme unité de quantité d'information (encore appelée « logon ») celle qui réduit l'incertitude de moitié ($n = N/2$), on a $k = 1/\lg 2$. Pour simplifier les formules, il est commode d'utiliser les logarithmes de base 2, ce qui donne finalement :

$$I = \lg_2(N/n)$$

[72] Quand $m^x = y$, on dit que x est le logarithme de y de base m .

Deux symboles (habituellement 0 et 1) sont suffisants pour représenter tous les nombres entiers. Un bit peut ainsi avoir une valeur 0 ou 1. À titre d'exemple, deux bits sont nécessaires pour représenter les nombres décimaux de 0 à 3 (0=00 ; 1=01 ; 2=10 ; 3=11); par contre 3 bits seraient nécessaires pour représenter les nombres décimaux de 0 à 7 (0=000 ; 1=001 ; 2=010 ; 3=011 ; 4=100 ; 5=101 ; 6=110 ; 7=111) etc.

La formule est que n bits permettent de représenter les nombres décimaux de 0 à $2^n - 1$. Un nombre quelconque représenté en binaire pondéré serait une succession de chiffres (bits) selon une pondération de poids (Tab.4). A titre d'exemple, le nombre décimal 58 s'écrirait en binaire « 111010 ».

Puisque l'on évoque à ce niveau le cas des systèmes binaires (numération à base 2), les exemples suivants illustreraient le fonctionnement du principe de codage selon les deux principes binaires numériques pondéré et réfléchi comme étant les plus utilisés dans l'automatisme et l'électromagnétisme actuel.

Poids pondérés						Valeur décimale
32	16	8	4	2	1	
0	0	0	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	0	2
0	0	0	0	1	1	3
0	0	0	1	0	0	4
0	0	0	1	0	1	5
0	0	0	1	1	0	6
0	0	0	1	1	1	7
0	0	1	0	0	0	8
0	0	1	0	0	1	9
0	0	1	0	1	0	10
...	...	...	...	...	...	...
1	1	1	0	1	0	58

Tableau 4 : La représentation binaire par système pondéré

Dans une table à valeurs réfléchies (Tab.5), le même exemple serait représenté différemment.

Poids réfléchis						Valeur
32	16	8	4	2	1	
0	0	0	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	2
0	0	0	0	1	0	3
0	0	0	1	1	0	4
0	0	0	1	1	1	5
0	0	0	1	0	1	6
0	0	0	1	0	0	7
0	0	1	1	0	0	8
0	0	1	1	0	1	9
0	0	1	1	1	1	10
...	...	...	...	...	...	...
1	0	0	0	1	0	58

Tableau 5 : La représentation binaire par système réfléchi

Ces deux méthodes de codage constituent aujourd'hui le soubassement des systèmes d'information automatisés. Seulement, si ces techniques de codage ont permis à l'homme de parler le langage de l'outil informatique, il n'est pas évident que ce dernier pourra réagir convenablement aux commandes de l'homme sans la présence d'un aiguilleur interne qui

analyse et agence les opérations de l'outil. Le traitement automatisé de l'information sous-entend évidemment une manipulation rationnelle et efficace de l'information. Ceci implique automatiquement une analyse formelle des processus de traitement. La résolution d'un problème concret ne peut s'accomplir par un ordinateur que si les opérations nécessaires à cette résolution peuvent être décomposées en un nombre fini d'étapes élémentaires. Chaque étape peut être traitée indépendamment de l'autre. En d'autres termes, on doit indiquer à l'ordinateur de façon précise et détaillée comment un problème donné doit être résolu. Un tel procédé de résolution est appelé *algorithme* que les programmes d'application eux même se chargent d'appliquer.

4. L'ALGORITHMIQUE

Par analyse algorithmique, il est communément admis qu'il s'agit d'un processus à travers lequel le système traitant effectue un choix d'un opérande quelconque en fonction d'une analyse logique elle-même basée sur l'analogie d'une série de variables qui lui sont proposées.

Si nous avons opté pour présenter le concept de l'algorithmique dans notre présent travail c'est pour mettre en exergue l'un des facteurs qui entrera en considération dans la configuration et le paramétrage des systèmes d'information multilingues où des structures itératives et alternatives entreront en action pour répondre à des besoins de traitement de l'information multilingue qui nécessitent à leurs tours un certain nombre de paramètres que le système saura, par alternance ou en simultané, mettre en place et déclencher à tout moment. L'origine orientale et particulièrement arabe du concept d'algorithmique aura, comme nous le verrons dans le chapitre réservé aux chiffres et à l'analyse des mécanismes d'incises, de tri et de bris textuels et chiffrés, une importance particulière dans la détermination de certains aspects d'usage des interfaces Homme-Machine et dans le traitement et l'accès à des données multilingues.

4.1. ETYMOLOGIE DU CONCEPT

Le concept d'algorithme est en fait inspiré du nom de son concepteur, le mathématicien arabe Al Khawarizmi (env. 820) dont le traité d'arithmétique *Kitab al-jabr wa al-muqabala* (« Règles pour restaurer et réduire », c'est-à-dire du passage d'une quantité d'un membre à l'autre d'une équation et de la réduction des termes semblables), a servi de source pour l'algèbre et les règles de calcul modernes basées sur la représentation décimale des nombres. Publié entre 813 et 830 à Bagdad, ce traité est considéré comme la première œuvre qui fait référence à une nouvelle discipline à part entière, en l'occurrence l'algèbre. « *Le terme d'algèbre apparaît dans un titre – al-jabr et al-muqabala y désignent à la fois une discipline et deux opérations ; soit, par exemple, $x^2 + c_2 - bx = d$, avec $c > d$; l'algèbre consiste à transposer les expressions soustractives ($x^2 + c = bx + d$), et al-muqabala à réduire les termes semblables : $x^2 + (c - d) = bx$* »^[73].

L'objectif d'al-Khwarizmi était d'élaborer une théorie des équations résolubles par radicaux, qui servirait de base pour résoudre de façon univoque des problèmes arithmétiques et géométriques nécessaires dans divers domaines comme le commerce, la topographie (arpentage des terres), etc. La popularité des mathématiques, et le rôle de la nouvelle science de l'algèbre considérée depuis le Xe et le XIe siècle par Al Farabi (872-950) et Avicenne (980-1037) comme discipline à part entière, ont permis de mettre au point ce qu'il est convenu d'appeler les mathématiques arabes ; résultat d'environ 7 siècles (IXe-XVIIe siècle) de labeurs de savants de toutes origines, de toutes

[73] Roshdi RACHED, Georges C. ANAWATI. Islam : les mathématiques et les autres sciences. In : Encyclopædia Universalis France S.A.1995. 12-706 b

religions, mais toujours restant en langue arabe. C'est justement au niveau de ce détail linguistique que vont se jouer beaucoup de considérations d'usage, d'ergonomie, d'adaptabilité et de conflit structurels et culturels des produits informatiques actuels dont nous évoquerons les plus remarquable à travers le noyau dur de cette étude : la fonction de calculabilité dans les algorithmes et la bidirectionnalité des textes multilingues.

4.2 DEFINITIONS DU CONCEPT

Quoique intuitive, la notion d'algorithme est difficile à définir formellement. Les diverses sources que nous évoquons ci-après convergent vers la définition qu'*un algorithme est une suite de règles, de raisonnements ou d'opérations, qui transforment des grandeurs données (données d'entrée) en d'autres grandeurs (données de sortie).*

Défini par l'Afnor comme « *ensemble des règles opératoires et des procédés, définis en vue d'obtenir un résultat déterminé, au moyen d'un nombre fini d'opérations* », l'algorithme est également présenté par Robert Reix comme un procédé de calcul et de traitement qui « *décrit une succession d'opérations à exécuter dans un certain ordre et sous certaines conditions pour passer des données de base aux résultats. Il correspond à un processus, découpé en étapes dont l'enchaînement permet d'arriver au but fixé* »^[74].

Selon la définition de Hebenstreit « *Un algorithme est une suite finie de règles à appliquer dans un ordre déterminé à un nombre fini de données pour arriver, en un nombre fini d'étapes, à un certain résultat, et cela indépendamment des données* »^[75].

Philippe Flajolet précise que « *L'objet de l'algorithmique est la conception, l'évaluation et l'optimisation des méthodes de calcul en mathématiques et en informatique. Un algorithme consiste en la spécification d'un schéma de calcul, sous forme d'une suite d'opérations élémentaires obéissant à un enchaînement déterminé* »^[76].

Du point de vue de l'informatique, le mathématicien Jacques RIGUET donne la définition suivante d'un algorithme: « *On appelle algorithme une méthode générale pour la résolution d'une classe de problèmes, fixée en tous ses détails par des règles dépourvues de sens, de façon à ce qu'on puisse l'appliquer sans avoir à la comprendre.* »^[77]

4.3 STRUCTURE ALGORITHMIQUE

Les algorithmes varient en complexité en fonction de la nature des problèmes réels qu'ils adressent. Cependant, tous sont construits autour d'un prototype unique conçu sous forme d'une combinaison des deux formes élémentaires de structures : des structures itératives et des structures alternatives.

[74] Robert REIX. Traitement des données. Paris : Foucher, 1984. p.189.

[75] Jacques HEBENSTREIT. Principes Informatiques : Algorithmes et décidabilité. Op. Cit.

[76] Philippe FLAJOLET. Algorithmique. In : Encyclopædia Universalis France S.A.1995. 1-814 b

[77] Cité par Daniel WEILER. L'algorithmique à l'URL : <http://w3.restena.lu/al/pascal/chaps/chap02/chap0202.htm>

4.3.1 Structures itératives

Par itération on sous-entend répétition. Les algorithmes sont en fait appelés à gérer des situations qui font appel de façon continue à des routines ou des fonctions répétitives afin d'arriver à un résultat escompté. Pour un tel fonctionnement, la notion de boucle, à l'opposé de l'analyse linéaire, est fondamentale pour réaliser un ensemble d'opérations répétitives toujours dans le même ordre un nombre fini de fois.

La boucle est en fait contrôlée par le mécanisme de validation qui met fin à l'itération conformément à une expression logique calculable par le système automate (ordinateur) prévue comme condition de fermeture de la boucle.

Dans la définition d'un algorithme, deux variantes d'itération peuvent être utilisées conformément à la manière dont est contrôlée la sortie de la boucle :

- La boucle « Tant que ... faire ... » : est réalisée selon la logique d'un filtrage pré-opérateur à travers la structure TANT QUE (l'expression logique est reconnue VRAI) FAIRE (un ensemble d'actions prédéfinies).

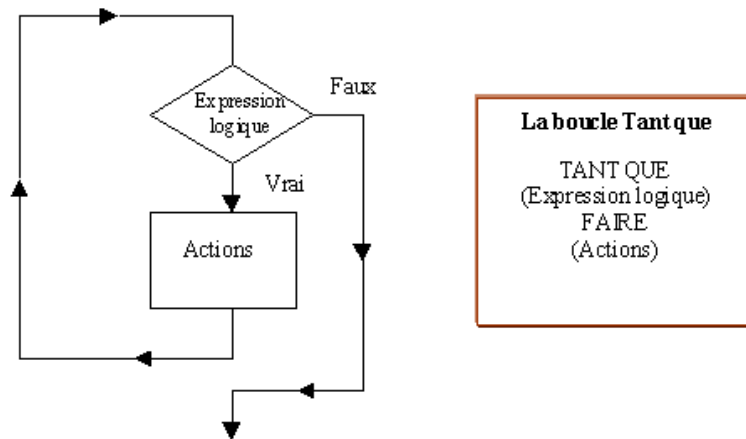


Figure 11 : Structure itérative à condition pré-opérateur

- La boucle « Répéter ... jusqu'à ... » est réalisée par contre conformément à la logique d'un filtrage post-opérateur en utilisant la structure REPETER (l'ensemble d'actions prédéfinies) JUSQU'A (ce que l'expression logique soit reconnue VRAI).

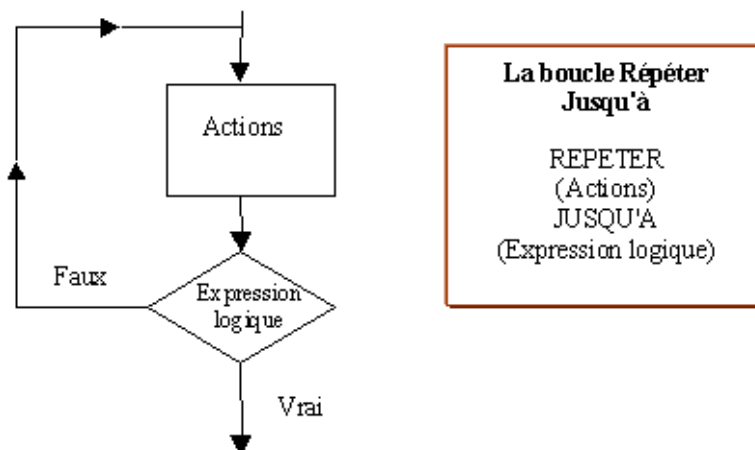


Figure 12 : Structure itérative à condition post-opérateur

4.3.2 Structures alternatives

La structure alternative correspond à une bifurcation dans le processus calculatoire de l'algorithme pour proposer un choix entre deux catégories d'opérations exclusives.

La structure générale d'une structure alternative est construite autour de l'expression de condition SI : SI (Expression logique) ALORS (exécuter un premier ensemble d'actions) SINON (exécuter un deuxième ensemble d'actions).

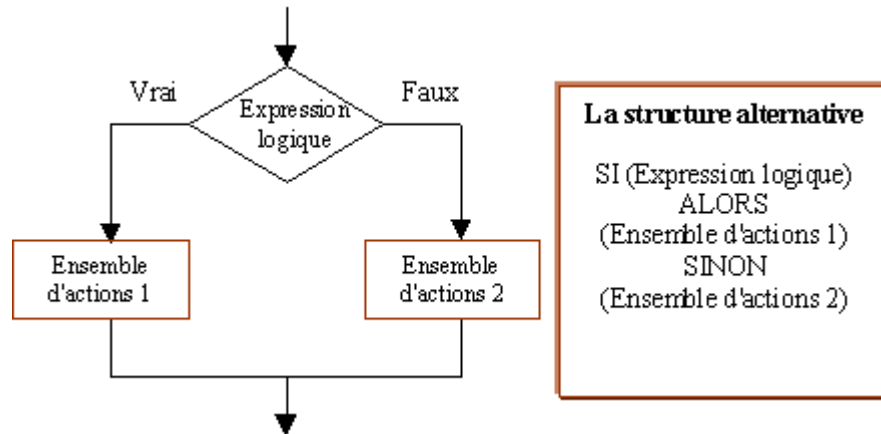


Figure 13 : Structure alternative

Cette forme simpliste de la structure alternative ne pose cependant pas de problèmes majeurs. Seulement, quand la proposition à tester est complexe (plusieurs conditions à tester, conditions imbriquées...) la structure alternative est soumise à des risques d'erreurs qui peuvent augmenter avec le nombre d'alternatives à combiner.

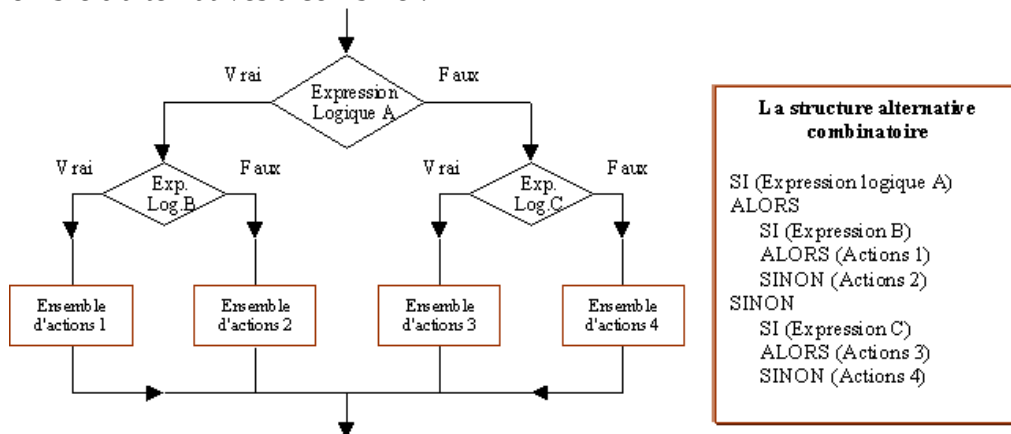


Figure 14 : Structure alternative combinée

L'une des solutions à apporter devant la complexité et le nombre élevé des structures alternatives pour la construction d'un algorithme est la définition de tables de décisions qui constituent un mode de représentation adapté aux algorithmes à structures alternatives combinées.

A Enoncé des conditions		C Entrée des conditions	RÈGLE	SI
B Enoncé des actions		D Entrée des actions	DÉCISION	ALORS EXECUTER
Souche		C o r p s		

Figure 15 : Structure générale d'une table de décision

Source : Robert Reix. *Traitement des données*. Paris : Foucher, 1984. p.218

Le fonctionnement d'une table de décision consiste à déclarer en bloc et au niveau des deux parties qui la constituent, la souche et le corps :

La souche

- ❑ A : une série de conditions pour décrire les expressions logiques définissant les alternatives;
- ❑ B : une série d'actions définissant les opérations à exécuter selon l'ordre de leur énonciation;

Le corps

- ❑ C : une entrée des valeurs des conditions définissant les expressions logiques;
- ❑ D : une entrée des actions où sont déclarées les valeurs précises des opérations à exécuter.

L'avantage de cette procédure de construction d'algorithmes est sans doute celle de permettre une combinaison assez souple de plusieurs valeurs de conditions combinées à l'aide de l'opérateur logique « ET » souvent appelée « Règle ». En parallèle, les règles dictent l'exécution d'une suite d'opérations dans une phase appelée « Décision » réduisant ainsi l'expression logique composée à une formulation du genre :

SI

Condition 1 ET condition 2 ET condition 3 ET ...

ALORS

Exécuter la décision (ensemble des opérations)

4.4 NATURE ET DOMAINES D'APPLICATION

La notion d'algorithmes a toujours existé depuis l'antiquité pour servir dans des domaines comme ceux de l'arithmétique ou de la géométrie. Les règles de calcul des distances et des surfaces des civilisations égyptienne et grecque ; l'algorithme d'Euclide (env. 300 av. J.-C.) pour le calcul du plus grand commun diviseur de deux nombres^[78] ; le schéma de calcul d'Archimède (~287-~212

[78] Voir l'ouvrage de J. ITARD. Les livres arithmétiques d'Euclide. Paris : Herman, 1962.

av. J.-C.)^[79] pour calculer le nombre π , les équations de Diophante d'Alexandrie (env.250 J.-C.)^[80] basées sur plusieurs méthodes de résolution d'équations en nombres entiers ... constituent des étapes de l'évolution du concept d'algorithmes à travers le temps.

Ensuite ont été étudiées les méthodes numériques de résolution d'équations algébriques (algorithme de Newton^[81], méthode d'élimination de Carl Friedrich Gauss^[82]), puis d'équations différentielles. Enfin, l'avènement de calculateurs électroniques, avec le début de la Seconde Guerre mondiale, a entraîné un renouvellement complet de l'algorithmique, appliquée désormais soit à des problèmes de taille bien supérieure à celle des problèmes traités manuellement (d'où la nécessité de méthodes nouvelles), soit surtout à des types de problèmes nouveaux, comme le tri, la recherche rapide d'informations non numériques (base de données) ou l'optimisation combinatoire.

En fait, l'algorithmique (science qui étudie l'application des algorithmes à l'informatique) est par définition « *la conception, l'évaluation et l'optimisation des méthodes de calcul en mathématiques et en informatique* »^[83] par la spécification d'un schéma de calcul conforme à une série de routines élémentaires qui correspondent à une logique d'enchaînement établie.

À travers toutes ces définitions, nous pouvons déduire qu'un algorithme agit sur une quantité d'information selon un procédé systématique complètement indépendant du caractère sémantique des données traitées. Son action principale est essentiellement orientée vers les différentes phases du traitement physique des données qui lui sont soumises. Daniel WEILER lui confère cinq règles de base qu'il est censé vérifier :

1) L'entrée des données (Input)

Pour fonctionner, un algorithme doit disposer d'un certain nombre de variables définies au préalable dans son énoncé et auxquelles il est censé s'appliquer. Ces données d'entrée sont des valeurs exactes définies avant son exécution.

2) Les éléments de sortie (Output)

En contre partie, un algorithme doit disposer dans son énoncé d'une ou de plusieurs données de sortie produites par l'exécution de l'algorithme en question sur la base des données d'entrée.

3) La terminaison

Un algorithme prévoit un cycle fini d'opérations doit se terminer après un nombre déterminé de pas.

[79] Voir : Archimède dans les mathématiques arabes. In : I. MUELLER. Dir., Essays around the mathematical sciences of Greeks. Apeiron, 1991.

[80] Voir Pierre de FERMAT. Précis des œuvres mathématiques et de l'arithmétique de Diophante. Ed. E. Brassinne, repr. En fac-sim. De l'éd. de Toulouse, 1853.

[81] Voir le traité d'Isaac NEWTON (1642-1727) « *Principes mathématiques de la philosophie naturelle* = Philosophiae naturalis Principia Mathematica), rééd. J.Gabay, 1990.

[82] Voir l'œuvre de G.M. RASSIAS. The mathematical Heritage of C.F. Gauss. World Scientific Publ., River Edge, 1991.

[83] Jacques HEBENSTREIT. Principes Informatiques : Algorithmes et décidabilité. In : Encyclopædia Universalis France S.A.1995. 12-306 b.

4) La clarté

L'application des règles et des procédures ainsi que l'ordre dans lequel elles s'enchaînent doivent observer une désambiguïté stricte. C'est à dire que l'algorithme décrit en détail la démarche qu'il doit suivre et les alternatives à toutes les situations qui se présentent. C'est la tâche reconnue sous le nom d'**affinement progressif** de l'algorithme ou **conception descendante**.

5) L'efficacité

L'efficacité d'un algorithme se mesure en fonction de deux critères essentiels :

- La simplicité des opérations exécutées en un temps fini ;
- Le nombre d'opérations fini et limité qui permet l'exécution et la terminaison d'un algorithme dans un temps raisonnable.

L'application des algorithmes dans un contexte d'automatisme s'était développée depuis 1936 avec la machine Turing^[84]. Il s'agit d'une structure qui s'appuyait sur la technique de segmentation de séquences finies d'opérations arithmétiques ou logiques écrites sous forme d'un programme. Cette technique allait révolutionner le monde de l'informatique par la mise en place d'une série de nouveaux modèles (langages) de programmes évolués^[85] qui substituent au langage ordinaire un autre langage qui ne soit pas simplement descriptif, mais plutôt à caractère opératoire tel le cas des langages algorithmiques. « *De tels langages ne sont réellement utilisables que si l'on effectue jusqu'au bout les opérations qu'ils requièrent. Et celles-ci peuvent être fort longues. Les machines viennent à point nommé pour rendre effectivement réalisable le passage du langage ordinaire aux langages de type algorithmique* »^[86].

L'algorithme de l'addition permet, à titre d'exemple, d'additionner deux nombres quelconques en progressant à partir des chiffres situés les plus à droite dans une évolution incrémentielle de droite à gauche. Partant de l'idée qu'un algorithme est une description du chaînage des opérations à réaliser, il est clair que la façon comment elles seraient rédigées, dépendra du dispositif (homme ou machine) chargé de le faire tourner. Un algorithme écrit en arabe par exemple n'est pas destiné pour celui qui ignore cette langue. Partant de cette constante, un programme destiné à un ordinateur est tout particulièrement un algorithme. Un langage de programmation est un langage permettant de rédiger un algorithme qu'un ordinateur saura exécuter.

[84] Les machines de Turing sont des machines abstraites conçues en 1936 par Alain Mathison Turing (1912-1954) sous forme d'une succession d'instructions capables d'agir de manière séquentielle sur des informations d'entrée pour fournir un résultat. C'est les débuts du formalisme du concept d'algorithme et de la théorie de calculabilité.

[85] Les générations des langages de programmation sont généralement classées en fonction des méthodes d'écriture des instructions soumises à la machine. Les programmes de première génération, en l'occurrence les langages de programmation (machines et assembleurs) de bas niveau, décrivent au détail près toutes les instructions que la machine est censée réaliser dans l'exécution d'une opération (i.e. les adresses des allocations mémoires). Chaque ligne de code constitue une action unique pour le système. La deuxième génération de langages, les langages évolués (i.e. Basic, Pascal, Cobol, Fortran, APL, PL/1...), nécessitent moins de détails de programmation bien qu'ils restent, à l'instar de leurs prédécesseurs des langages procéduraux (méthodes d'apprentissage des systèmes pour la réalisation étape par étape des différentes procédures du programme). Une troisième catégorie de programmes d'un plus haut niveau est identifiée comme non procédurale. À titre d'exemple, la somme d'une série de valeurs numérique sur Excel se limite à une activation de la procédure prédéfinie suivie d'une sélection des cellules à additionner. L'opération ne nécessite plus un apprentissage répétitif du système sur les modalités d'addition des valeurs numériques.

[86] Jean LADRIERE. Cybernétique. In : Encyclopædia Universalis France S.A.1995.6-1001 a.

5. CONCLUSION

Nous avons essayé à travers ce chapitre de mettre en évidence la présence des phénomènes parallèles à toute activité ou tout fonctionnement d'un système dont les soubassements résultent de la confluence historique ou immédiate de plusieurs facteurs catalysants. Les interfaces Homme-Machine, comme le précise Harold Thimbleby, font intervenir deux disciplines dont les assises culturelles sont complètement différentes : la psychologie et les sciences informatiques. Il décrit cette distinction dans les termes suivants : « *Psychology is concerned with people ; computer science with computer machinery. Psychologists are supposedly sympathetic and understanding ; computer scientists are supposedly mathematical and precise. Psychologists have enough trouble understanding people even when they are not using computers ; computer scientists have enough trouble getting programs to work en when they are not being used by people* »^[87]. Or ces deux disciplines, plus particulièrement impliquées dans les deux processus de conception et d'évaluation des interfaces Homme-Machine, constituent les deux piliers de base du contexte de conception et d'évaluation des IHMs. Au delà de ces facteurs directs, il y a eu comme nous l'avons évoqué tout au long de ce chapitre une évolution lente et une adaptation progressive de plusieurs autres phénomènes psychotechniques qui ont contribué à la genèse d'un terrain favorable au développement des interfaces Homme-Machine. Nous synthétisons cette confluence interactive, directe et indirecte, historique et immédiate à travers l'illustration suivante :



Figure 16 : Soubassements de l'Interactivité Homme-Machine

[87] Harold THIMBLEY. User interface Design. New York : ACM Press, 1955. pp.6-7.

CHAPITRE TROISIÈME

CODAGE, NORMALISATION ET TRANSPARENCE LINGUISTIQUE

Le principe du codage de l'information, comme nous l'avons présenté dans les chapitres précédents, visait (dans sa forme graphique) la représentation des caractères dans une autre forme que celle du transcrit et du parlé. Pour transmettre une donnée, il faut donc associer à chaque symbole à transmettre une suite unique d'éléments binaires et vice versa. Le nombre d'éléments binaires qui identifieront les symboles à transmettre est relatif au nombre de symboles que l'on veut transmettre.

L'extension des champ d'application de ces mécanismes de représentation et des techniques de codage et de décodage des messages au delà des paires d'émetteurs et récepteurs isolées pour toucher tous les membres d'une même communauté, a abouti à l'expression d'un besoin de normalisation (ou plutôt de standardisation!) formulé depuis le début de ce siècle pour tirer profit des nouvelles technologies naissantes. Le code morse et les cartes perforées de Hollerith (1860-1929) puis un peu plus tard le modèle privé BCD (Binary Coded Decimal) proposé par IBM qui donna lieu au standard EBCDIC (Extended Binary Coded Decimal Interchange Code)^[1], étaient des étapes stratégiques dans ce processus de codification des données qui visait la transparence linguistique dans les nouvelles procédures de traitement et d'échange automatisés de l'information.

Le premier jeu de caractères codés était un code à 5 bits patenté par Jean-Maurice-Émile Baudot (1845-1903) en 1874 pour faire fonctionner à l'énergie électrique l'ancêtre des téléimprimantes qu'il avait mis au point.

En 1932 le CCITT (Comité Consultatif International Télégraphique et Téléphonique) a adopté le code Baudot à 5 bits pour en faire le code du réseau télex international encore en vigueur sous le nom l'Alphabet International Télégraphique n°2 ou code CCITT n°2 ou simplement le code Baudot^[2]. Ce standard a été adopté depuis lors comme premier modèle binaire^[3] ou combinaison binaire normalisée^[4] destinée au fil du temps à servir dans des champs d'application assez porteurs : le stockage et la transmission des données).

Étant donné le nombre limité de la combinaison binaire à 5 bits qui ne peuvent engendrer, selon la technique de pondération expliquée antérieurement, que trente-deux combinaisons possibles ($2^5 = 2*2*2*2*2=32$), insuffisantes pour les 26 lettres de l'alphabet et les 10 chiffres, les téléimprimantes fonctionnant sur le modèle Baudot utilisent la technique des jeux de caractères décalés « Shifed Coding Sets » à la manière d'une machine dactylographique traditionnelle. Cette technique aurait une grande répercussion sur les modèles de codes futurs (nous en verrons le cas de la norme ISO/IEC 2022) utilisant la technique d'extensions de codes de caractères. Le code

[1] EBCDIC est une norme posée par IBM depuis les années 1960 sur le modèle des cartes perforées de Hollerith. C'était au début une norme de codification binaire sur 4 bits en mode décimal (Binary Coded Decimal) d'où son nom d'origine BCD. C'était une norme à poids variables de sorte que l'on avait au début plusieurs codes BCD. C'est avec l'apparition de la valeur d'« Octet » qu'IBM a procédé en 1965 à une extension de son code BCD pour devenir un code à 6 puis à 7 et 8 bits où les variantes nationales allaient avoir une place à côté de la série des codes internationaux.

[2] Il a été réédité en 1988 comme recommandation CCITT S.1

[3] Les modèles binaires ont en fait joué un rôle essentiel depuis l'apparition des ordinateurs pour stocker les données dans un langage machine structuré en combinaisons binaires de longueur fixes. Les plus utilisés étaient les mots de 24, 32, 36, 48, 60, 64 bits.

[4] Ils ne comprenaient à ce temps là que les lettres majuscules, les chiffres, quelques signes de ponctuation et certains codes réservés.

Baudot à 5 bits permettrait ainsi la représentation de 58 signaux différents (26 signaux pour chaque mode plus 6 non affectés)

Il était cependant admis dès le début que l'internationalisation de ce modèle passerait obligatoirement par son adaptation aux besoins de certains pays et certaines langues nationales. Malgré les modifications de 1932 apportées au standard Baudot et l'extension de son modèle binaire à 6 bits, trois positions uniques ont été allouées pour répondre à des usages nationaux.

Plus tard, l'histoire sera pourtant féconde d'un grand nombre de normes et de standards qui ont été proposés ou « *redéfinies pour coder les caractères avec plus de bits ou moments pour s'adapter à diverses langues, dont les célèbres ISO/IEC-8859-n sur 8 bits et Unicode à 16 bits, étendu à 32 bits pour devenir ISO/IEC-10646.* »^[5] (Voir [annexe A](#)).

Nom du code	Nombre de bits ou moments	Nombre de caractères
Télex	6	64
ASCII	7	128
ISO-Latin 1	8	256
UNICODE	16	65536
ISO/IEC 1646	32	> 2 milliards

Tableau 6: Principales normes de codage.

Source : Jacques André et Michel Goossens. *Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646.* In : Cahiers GUTenberg n°20-mai 1995.pp.1-53.

1. L'ASCII : L'HISTOIRE D'UN STANDARD DEvenu UN MODELE INTERNATIONAL

En effet, à partir des années 50, l'usage croissant de méthodes électroniques a nécessité l'adoption de standards qui devraient servir des domaines à échelle mondiale où l'échange des données était d'importance fondamentale non pas aux États Unis seulement mais aussi aux autres puissances émergentes : l'Europe et le Japon. C'était donc sous l'emprise d'un besoin urgent d'échange et de transfert de données que plusieurs normes ont été définies dans ce sens pour répondre aux exigences techniques et culturelles des grands opérateurs du monde de la transmission des données. Depuis l'aspect privé et propriétaire de l'EBCDIC, les Américains ont œuvré en collaboration avec les instances internationales de normalisation pour développer un premier standard national : l'ASCII (American Standard Coded Information Interchange) ensuite normalisé et adopté en 1963 par l'Organisation Internationale de Normalisation (ISO) et le CCITT sous le nom de norme ISO 646.

L'ISO, créée en 1947, a en fait confié cette tâche au sous comité 2 de son comité technique 97 (ISO/TC97/SC2) qui œuvrait en parallèle avec les organisations américaines de normalisation, en l'occurrence la l'E.I.A. (Electronic Industries Association) puis l'A.S.A. (American Standards Association) qui deviendrait au fil du temps l'U.S.A.S.I. (United States of American Standards Institute) puis définitivement l'actuel A.N.S.I. (American National Standards Institute). L'A.N.S.I. était désigné à travers le sous comité X3.2 pour mettre en place une forme d'un code normalisé capable de représenter les données sur tous les médias standards de l'époque en particulier les bandes magnétiques et les cartes perforées. L'ASCII est né ainsi d'un effort de recherche américain afin de constituer un code de représentation de caractères avec lequel toutes

[5] Jacques ANDRE et Michel GOOSSENS. *Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646.* In : Cahiers Gutenberg n°20-mai 1995.pp.1-53.

les autres normes assureront toute la compatibilité. « D'ou la très grande importance de cette norme même si maintenant elle a vieilli »^[6].

1.1 LA « BIOGRAPHIE D'UN DINOSAURE »

Durant les premières périodes du développement du code ASCII, la tendance a été de suivre le modèle BCD pour définir le nombre de caractères à adopter. Le choix définitif de l'ISO/TC97/SC2 s'est inspiré de l'expérience de plusieurs normes de l'époque pour proposer une version commune de l'ASCII standard^[7]. Une première analyse comparative de ces normes de l'époque a donné comme modèle standard un code composé de 10 caractères numériques, 26 caractères graphiques et entre 11 et 25 caractères spéciaux, soit un modèle entre 46 et 64 caractères graphiques selon les estimations de l'époque pour l'échange des données^[8].

À travers l'analyse des mêmes normes de l'époque, il s'est également avéré qu'un standard correspondant aux besoins du contrôle des transmissions des données nécessitait une moyenne de 10 caractères de contrôle^[9]. Ceci laissait entrevoir que les codes de caractères combinés entre les caractères graphiques et les codes de contrôle seraient supérieurs aux présumés 64 possibilités de codes proposées au début.

Une première tendance vers la technique des caractères décalables (Shifted codes)^[10] a été proposée mais rapidement rejetée, bien qu'économique sur le plan du rapport entre le nombre de caractères et le modèle binaire concerné. Parmi les grands avantages des caractères décalables c'est qu'ils permettent le codage de plus de caractères que la taille du code binaire ne le permet grâce à la formule suivante :

$$2^{x+1} - y$$

où « x » représente le nombre de bits dans le code utilisé et « y » est le nombre de caractères décalables. Selon la même formule, la norme CCITT n°2 à 5 bits décalés propose 58 caractères différents dont 6 caractères non affectés.

$$2^{5+1} - y = 2^6 - y = 58 \rightarrow y = 2^6 - 58 = 64 - 58 = 6$$

Un code à 6 bits en permettra selon le même principe 111 dont 17 ne sont pas décalés. Cette méthode a été cependant rapidement rejetée car elle engendrait un grand taux d'entropie due à la non fiabilité des lignes de télécommunications de l'époque. Comme suite au phénomène de « perturbation »^[11] (Hit) très courant en télécommunications d'antan, un bit de valeur 1 est automatiquement converti en bit 0 et vice versa. Si le modèle binaire 11101 de la lettre Q est bruité pour devenir 11001 (l'équivalent de W), le mot LEQUEL deviendrait LEWUEL. Ce qui donne un facteur d'entropie limité aisément déductible à travers le contexte. Le cas où le « Hit »

[6] Jacques ANDRE et Michel GOOSSENS. Codage des caractères et multilinguisme. Op.Cit. p.9.

[7] Le CCITT n°2 disposait de 50 caractères dont 26 alphabétiques, 10 numériques, 3 emplacements de codes pour un usage national et 11 caractères spéciaux. Le code des cartes perforées de l'époque donnaient la possibilité de représenter 52 caractères graphiques dont 26 alphabétiques, 10 numériques, 16 caractères spéciaux. Le code Fieldata développé par l'armée pour les télécommunications disposait de 10 caractères numériques, 26 alphabétiques et 19 caractères spéciaux. Le code BCDIC disposait de 10 caractères numériques, 26 alphabétiques et 32 caractères spéciaux.

[8] Charles E. MACKENZIE. Coded Character Sets, History and Development. London : Addison-Wesley Publishing Company, 1980. p.213

[9] Le CCITT n°2 disposait de 7 caractères de contrôle, Fieldata en avait 9 et le BCDIC en comptait 7.

[10] Une extension de codes obtenue par combinaison des commandes de codes normaux (Shifted In) et spéciaux (Shifted Out).

[11] Des perturbations momentanées peuvent engendrer la mutilation des caractères transmis sur des lignes de télécommunication.

toucherait un grand nombre de caractères à la fois ou affecterait le code d'un chiffre^[12] ou celui d'un caractère décalable, le résultat est un message incompréhensible où la redondance est maximale et l'entropie est élevée.

La polémique sur le nombre de bits à adopter pour coder les caractères restait pour un temps très fluctuant entre l'économie des caractères décalables et la confusion des messages transmis. La fiabilité de l'information l'a finalement emporté^[13] et la tendance évoluait progressivement vers le code à 7 éléments binaires bien qu'un modèle à 8 éléments binaires était largement soutenu au départ. Le modèle 7 bits, qui allait garantir la représentation de 128 caractères indépendants, l'a emporté suite à un vote tenu par le sous comité X3.2 et argumenté par les faits suivants :

- ❑ Les 128 combinaisons d'un code à 7 bits étaient largement suffisantes pour les besoins d'échange et de contrôle des données entre la majorité des utilisateurs ;
- ❑ L'usage d'un code à 8 bits qui permet la mise en place de 256 combinaisons imposerait la contrainte d'enregistrer et de transmettre sur une base de 8 bits alors que la majorité des appareils d'entrée et de sortie de l'époque étaient des composantes à 7 bits;
- ❑ Un bit de parité est largement utilisé dans les opérations de lecture et d'écriture ainsi que pour le contrôle de parité dans les opérations de transmission de données. Le principe du codage à 8 bits (7 codes de caractères et un bit de parité) était le maximum possible pour l'enregistrement sur les supports de l'époque.

La décision a été définitivement prise en 1963 pour un codage à 7 éléments binaires dans une matrice de 16 lignes sur 8 colonnes ([Tab.7](#)). Cette matrice allait ensuite servir, dans ses deux formes standards (7 bits) ou étendues (8 bits) de support officiel pour toutes les représentations codées des autres langues. Une description détaillée du développement de ces normes est trouvée dans le livre de Mackenzie^[14].

1.2 LA STRUCTURE BINAIRE DE L'ASCII

La structure binaire de l'ASCII était une décision administrative plutôt que technique car elle allait adresser le problème de la présentation du jeu de caractères codés dans sa version visuelle.

Un jeu de caractères codés est par définition « *un ensemble de règles non-ambiguës qui définissent un jeu de caractères et établissent une correspondance biunivoque entre chaque caractère du jeu et sa représentation codée* »^[15]. Ces représentations codées sont traduites sous forme d'un tableau (matrice) qui montre les caractères alloués à chaque combinaison binaire d'un code.

Les codes en question couvrent deux sortes de caractères : des caractères graphiques qui servent à la transmission des données et des caractères de contrôle ([Voir Annexe B](#)) qui servent à exercer certaines actions sur l'ordinateur et ses périphériques.

Le comité a opté dans ce sens vers une matrice de 8 colonnes représentant les 3 bits ayant le plus grand poids du code (000, 001,...,111) et 16 lignes représentant les 4 bits d'un poids moindre

[12] Les télégraphistes d'antan retapaient tous les chiffres en séquentiel à la fin du message, bien que cette partie du message puisse aussi être « bruitée », comme garantie d'authenticité des chiffres dans le corps du message.

[13] Paradoxalement, deux caractères décalables ont été insérés beaucoup plus tard dans le code ACII dans les emplacement 14/0 et 15/0 (Shifted Out et Shifted In).

[14] Charles E. MACKENZIE. Coded Character Sets, History and Development. Op. Cit. 513 p.

[15] Norme Internationale : ISO 8859-6, Traitement de l'information - Jeux de caractères graphiques codés sur un seul octet. Première édition, 1987-08-15. p.1.

dans le code en question (0000, 0001,...,1111). La constitution d'un code à 7 bits serait ainsi conforme à la disposition suivante :

Élément	b7	b6	b5	B4	b3	B2	b1
Poids	64	32	16	8	4	2	1

Tableau 7 : Positions et poids des combinaisons binaires à 7 bits.

Cette disposition normalisée permet d'identifier les combinaisons binaires selon la notation x/y où « x » est le nombre de la rangée entre 0 et 7 dans la matrice globale (b7, b6, b5) et « y » est le nombre de la ligne entre 0 et 15 (b4, b3, b2, b1) (Tab.7).

Une décision a été prise pour réserver les 32 premières combinaisons binaires pour les caractères de contrôle. Cette disposition inclut le caractère NULL (0/0) et exclut le caractère DELETE (7/15) laissant place à 95 combinaisons binaires pour les caractères graphiques. La disposition des caractères graphiques (le caractère ESPACE inclus) était conçue dans un ordre utile pour des opérations de tri. Le caractère ESPACE est normalement trié avant tout autre caractère graphique ; raison pour laquelle il était placé en première position de la liste des caractères graphiques qui détiennent encore les 94 positions entre 2/0 et 7/14 avec des codes respectifs entre « 1100000 » (ESPACE) et « 1111111 » (DELETE). Cette distribution des positions des codes (32 caractères de contrôle introduits par NULL et 94 caractères graphiques encadrés par ESPACE et DELETE a dominé la modélisation des structures des jeux des caractères codés jusqu'à nos jours^[16] (Tab.7)

La matrice suivante (Tab.8) reproduit ces caractéristiques de la disposition des bits conformément à la norme actuellement en vigueur.

					B7	0	0	0	0	1	1	1	1				
					B6	0	0	1	1	0	0	1	1				
					B5	0	1	0	1	0	1	0	1				
						0	1	2	3	4	5	6	7				
B4	b3	B2	b1		0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	0	0	0	0	0	0	0	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	0	2	0	0	1	0	STX	DC2	"	2	B	R	b	r
0	0	1	1	0	3	0	0	1	1	ETX	DC3	#	3	C	S	c	s
0	1	0	0	0	4	0	1	0	0	EOT	DC4	\$	4	D	T	d	t
0	1	0	1	0	5	0	1	0	1	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	0	6	0	1	1	0	ACK	SYN	&	6	F	V	f	v
0	1	1	1	0	7	0	1	1	1	BEL	ETB	'	7	G	W	g	w
1	0	0	0	0	8	1	0	0	0	BS	CAN	)	8	H	X	h	x
1	0	0	1	0	9	1	0	0	1	HT	EM	(	9	I	Y	i	y
1	0	1	0	0	10	1	0	1	0	LF	SUB	*	:	J	Z	j	z
1	0	1	1	0	11	1	0	1	1	VT	ESC	+	;	K	[	k	{
1	1	0	0	0	12	1	1	0	0	FF	IS4	,		L	\	l	
1	1	0	1	0	13	1	1	0	1	CR	IS3	-	=	M	]	m	}
1	1	1	0	0	14	1	1	1	0	SO	IS2	.		N	^	n	~
1	1	1	1	0	15	1	1	1	1	SI	IS1	/	?	O	_	p	DEL

Tableau 8 : La table ASCII = ISO 646

[16] Il faut savoir que le code ASCII, dans sa première version standardisée, publiée en 1963, disposait uniquement des lettres capitales et que l'ASCII standard que nous connaissons aujourd'hui date de l'année 1968.

Par convention, une norme internationale attribue au moins un nom à chaque caractère en plus de spécifier un acronyme pour chaque caractère de contrôle y compris le caractère ESPACE (2/0) ainsi qu'un symbole pour chaque caractère graphique. Par convention aussi, seules les lettres capitales, les apostrophes et les tirets sont autorisés pour transcrire les noms des caractères. Pour les acronymes, seuls les lettres capitales et les chiffres sont d'usage. Les noms choisis afin de décrire les caractères graphiques reflètent leurs sens courants sans pour autant les restreindre ou leur imposer un sens particulier. Ils ne spécifient non plus, et dans aucun cas, un style ou une forme de police quelconque aux symboles graphiques qui représentent les caractères ([Voir Annexe C](#)).

Le principe de l'ASCII (ISO 646) est donc de fournir des codes (en appliquant la pondération adéquate) pour 128 caractères différents répartis comme suit :

- ❑ Les combinaisons binaires de 0/0 à 1/15 représentent 32 caractères de contrôle ;
- ❑ La combinaison binaire 2/0 représente le caractère ESPACE interprété comme caractère de contrôle et caractère graphique à la fois ;
- ❑ Les combinaisons binaires de 2/1 à 7/14 représentent 86 caractères graphiques ;
- ❑ La combinaison binaire 7/15 représente le caractère de contrôle DELETE.

Ainsi cette table comprend en fait l'ensemble des caractères suivants :

- ❑ 34 caractères de contrôle y compris les caractères ESPACE et DELETE ([Voir Annexe B](#));
- ❑ 94 caractères graphiques visibles sur écran ou sur imprimante dont :
 - 82 caractères obligatoires ;
 - 52 lettres : A-Z et a-z (Majuscules et minuscules) ;
 - 10 chiffres : 0-9 ;
 - 20 signes de ponctuation ou autres : ! " % & ' () * + , - . / : ; < = > ? _ ;
 - 2 caractères au choix^[17] :
 - # ou £ ;
 - \$ ou ;₹
 - 10 positions réservées à des caractères d'usage national ([Tab.9](#)).

Ainsi, l'ISO 646 (1^{ère} éd. 1973, 2^e éd. 1983) spécifie des codes 7-bits pour 128 codes possibles dont 33 sont pour des caractères de contrôle, un pour l'ESPACE, et 94 pour les caractères graphiques. De ce dernier groupe, 82 sont uniques, 12 sont laissés ouverts pour la définition d'une version nationale d'ISO 646. Une Version de la Référence Internationale (IRV) est fournie sans aucune préférence nationale.

[17] André et GOOSSENS rapportent que ce « choix » a été imposé par certains états comme la Russie et la Grande Bretagne lors de la guerre froide pour contrecarrer la supériorité du Dollar. Ce dernier pourtant regagnera sa position au dépend du symbole monétaire international « ₹ » à partir de la période de la Perestroïka. Source : Jacques André et Michel Goossens. Codage des caractères et multilinguisme. Op.Cit. p.11.

1.3 L'IRV, L'EMBRYON MULTILINGUE DE L'ASCII

Selon Johan W. Van Wingen^[18], l'ASCII a été conçu, dans sa structure définitive, pour bien servir deux premières fonctions essentielles : le stockage et la transmission des données. D'ailleurs, la présence des caractères NULL (0/0) et DELETE (7/15), codés par « 0000000 » et « 1111111 » respectivement, justifient l'influence du traitement et de l'échange des données par cartes perforées^[19].

Par contre, pour le traitement interne par un ordinateur, l'ASCII n'était pas très bien adapté. Un mot binaire de 7 bits était à peine utilisable. Pour la représentation interne, des codes de 6 ou 8 bits étaient beaucoup mieux adaptés étant donné qu'un « octet » de 6 bits pourrait être contenu 4, 6, 8 ou 10 fois dans des mots machines de 24, 36, 48 ou 60 bits respectivement. Un « octet » de 8 bits pourrait à son tour être contenu dans un mot de 32 ou 64 bits. Il n'est donc pas surprenant de voir beaucoup de fabricants d'ordinateurs définir leurs propres jeux de caractères sur un codage à 6 ou 8 bits pour des usages appropriés spécifiques.

En fait c'est depuis les débuts des années 1970 que les utilisateurs finals ont éprouvé un grand nombre de problèmes avec les jeux de caractères et les codes ASCII et EBCDIC à la fois.

Il était prévu que depuis son apparition, l'ASCII allait, à l'instar du code Baudot, être revendiqué à échelle internationale. Parce que des pays autres que les USA ont des exigences différentes au contenu du jeu des caractères codés, le document référencé ISO R 646 contient « *des variantes nationales (parfois plusieurs pour un même pays - c'est le cas de la France)* »^[20] et une version internationale de référence (IRV) mal adaptée et donc mal appliquée où les « *positions optionnelles sont affectées d'un caractère précis* ». Une fois appliqué, le résultat est une Version Nationale, aux mêmes termes que l'ASCII qui reste la Version Nationale Américaine (Tab.8). En effet, ASCII est simplement la Version Nationale Américaine d'ISO 646. Elle diffère de l'IRV simplement pour avoir un symbole \$ au lieu du signe de monnaie. Mais l'allemand, le suédois, le danois/norvégien et toutes les autres langues remplacent les lettres accentuées au niveau de la plupart des 12 emplacements. Qu'est-ce que cela veut dire ? Si l'on envoie des crochets en Norvège, ils arrivent comme Æ et Å (A-bague). Cette pratique met une barrière entre les deux mondes anglophone et non anglophone causée par le nombre des caractères limités à 94. Le principe de correspondance du code à caractère unique avait été ainsi abandonné.

• ISO-IR 2	International Reference Version de l'ISO 646:1983
• ISO-IR 4	Version britannique
• ISO-IR 6	International Reference Version de l'ISO/IEC 646:1991 (également version américaine)
• ISO-IR 10	Version suédoise
• ISO-IR 11	Version suédoise pour les noms
• ISO-IR 14	Version japonaise pour les caractères romains

[18] Johan W. Van WINGEN. Coded character sets and programming languages. *Leiden, the Netherlands* : ISO/IEC JTC1/SC2 N 1961R, ISO/IEC JTC1/SC22 N 578R, September 1988. Revised April 1989. Version 1.2

[19] L'effet des cartes perforées est visible à ce niveau. Un bit de valeur « 1 » était représenté par une perforation et un bit « 0 » par l'absence d'une perforation. Quand une rangée de 7 bits de valeur « 0 » serait traitée, le code « 0000000 » représenterait un caractère NULL (absence de tout effet). Par contre étant donné que les perforations ne peuvent pas être annulées, un caractère erroné est converti en « 1111111 » symbolisant le caractère DELETE. Une fois traité ou transmis, un autre caractère peut toujours le substituer.

[20] « Il y avait une grande incohérence d'un pays francophone à l'autre à tel point d'ailleurs que la France a abandonné cette norme z62010 au profit de l'ASCII en 1983. » Source : Jacques ANDRE et Michel GOOSSENS. Codage des caractères et multilinguisme. Op.Cit. p.12.

• ISO-IR 15	Version italienne
• ISO-IR 16	Version portugaise, ECMA (Olivetti)
• ISO-IR 17	Version espagnole, ECMA (Olivetti)
• ISO-IR 21	Version allemande
• ISO-IR 60	Version norvégienne
• ISO-IR 69	Version française
• ISO-IR 84	Version portugaise, ECMA (IBM)
• ISO-IR 85	Version espagnole, ECMA (IBM)
• ISO-IR 86	Version hongroise
• ISO-IR 141	Version serbo-croate et slovène
• ISO-IR 170	Caractères invariables de l'ISO/IEC 646 (82 caractères)

Tableau 9 : Les Références Nationales ISO/IEC 646

Avec les règles de l'ISO R 646 datée en 1968 (révisées en 1973 et en 1983)^[21] il est devenu possible de coder des textes en danois ou en suédois qui comportent un alphabet de 29 lettres au prix de perdre 6 caractères spéciaux. Des besoins supplémentaires tels que les caractères accentués (i.e. pour le français) ou les caractères spéciaux supplémentaires ne pourraient cependant pas être satisfaits.

La version la plus récente de cette norme est la troisième édition de 1991, (ISO/IEC 646:1991), qui a remplacé celle de 1983. Dans ces versions successives, aux 10 positions réservées pour les usages nationaux, deux positions ont été rajoutées pour inclure les deux paires alternatives : le signe de nombre (#) versus le signe monétaire du Livre Sterling (£) d'une part, et le signe dollar (\$) versus le signe monétaire générique (¤) d'autre part (Tab.10).

Version de référence	#	¤	@	[	\	]	^	'	{		}	~
Allemagne (DIN66003)	#	\$	§	Ä	Ö	Ü	^	'	ä	ö	ü	ß
Belgique	#	\$	à	°	ç	§	^	'	é	ij	è	~
Espagne	#	\$	.	i	Ñ	Ç	¿	'	'	ñ	ç	"
France (NF Z6201 0/1 982)	£	\$	à	°	ç	§	^	μ	é	ù	è	"
Grande Bretagne	£	\$	@	[	\	]	^	'	{		}	~
Suisse Romande			à		ç				é	ù	è	~
USA (Norme ASCII)	#	\$	@	[	\	]		'	{		}	~

Tableau 10 : Les caractères optionnels de la version de référence IRV de la norme ISO 646 classés par variantes nationales.

Source : Jacques André et Michel Goossens. Codage des caractères et multilinguisme. Op.Cit. p.11.

Les codes à 7 bits qui ont succédé à l'ASCII, pour couvrir des langues comme le grec, l'arabe, l'hébreu ... ont adopté le modèle de la structure de base de l'ASCII. Ils ont gardé les 32 caractères de contrôles situés entre les positions 0/0 et 1/15, le caractère ESPACE (2/0) et DELETE (7/15) tout en changeant les 94 caractères graphiques qui restent situés entre 2/1 et 7/14. Un petit problème reste quand même à résoudre : quand un codage à 7 bits est utilisé dans un

[21] En 1988, ISO 646 a repris exactement le codage ASCII (Version Nationale Américaine) ce qui justifie la présence de caractères comme & \$ £ ¢ @ # ... sur les claviers des ordinateurs d'aujourd'hui.

environnement à 8 bits (le cas de la majorité des ordinateurs), le bit le plus important (b8) prend automatiquement la valeur 0. Ceci donne au caractère NULL la combinaison binaire « 00000000 » alors que le caractère DELETE devient « 01111111 » ce qui contredit la règle initiale de l'ASCII du « tout zéro » pour le caractère NULL et du « tout un » pour le caractère DELETE.

Avec l'apparition des micro-ordinateurs PC et leurs systèmes d'exploitation, surtout l'IBM PC et son système MS-DOS, le nouveau concept de l'internationalisation est apparu avec le mécanisme des « pages de codes ». Il s'agit d'une série de jeux de 256 caractères utilisables dans différentes parties du monde. Les termes « jeux de caractères » et « Pages de codes » sont généralement appréhendés comme deux synonymes dont le rôle essentiel est d'adapter les opérations d'entrée (claviers) et de sortie (écrans) de tout système informatique aux besoins de l'utilisateur. Des codes pages pour chaque langue sont actuellement disponibles pour les plates-formes PC et gros systèmes et c'est à l'utilisateur d'accomplir le paramétrage de son contexte de travail en fonction de son besoin linguistique (Voir [annexe D](#)).

2. L'EXTENSION DES CODES : UN TREMPLIN VERS UN CONTEXTE MULTILINGUE DE CODAGE ET D'ECHANGE DE L'INFORMATION

Conduit par cet élan de technicité qui a marqué les débuts de la révolution informatique, tout le contexte informationnel universel s'est ancré dans l'identité des origines. Le codage ASCII, accompli à l'intérieur des ordinateurs et transitant entre les réseaux téléinformatiques, ne tenait compte que d'un nombre limité de caractères. Ce *pêché originel* américain qui marque encore toutes les cultures et les langues surtout non latines au niveau de leur épanouissement universel à travers les technologies de l'information, explique cette hégémonie consciente persistante et protégée de la culture américaine et de ses modes d'action. Car, comme le précise Bernard Cassen, « *c'est dans la moindre volonté hégémonique consciente, tant l'évidence de l'universalité de leur langue est intériorisée par les Américains, qu'a été édifié un réseau planétaire qui ne fonctionne pleinement qu'en anglais* »^[22].

Cette alliance support/logiciels/contenu allait ensuite émerger sous forme de standards qui élèveraient la langue anglaise au statut de *lingua franca* ou *langue par défaut* de tous les systèmes d'information et de communication. L'ISO^[23], l'IEC^[24], l'ECMA^[25], le CCITT^[26], l'IEEE^[27] ... sont autant d'institutions internationales qui sont parties dans leur longue marche de normalisation sur le dénominateur commun de la langue anglaise.

Or l'internationalisation de l'outil et sa pénétration dans des contextes linguistiques divers (90% du reste de la population mondiale) a engendré des problèmes d'usage imposés par les limites observées au niveau de l'ASCII dans la représentation des caractères particuliers aux autres langues du monde. Ceci a engendré un recours à une large panoplie de jeux de caractères nationaux complètement disfonctionnels dans une optique d'échange et de partage vu leur incompatibilité les uns avec les autres.

[22] Bernard CASSEN. Le Tout-anglais n'est pas une fatalité. Le Monde Diplomatique : Manière de voir, octobre 1996, pp.81-82.

[23] International Standards Organisation. <http://www.iso.ch/welcomef.html>

[24] International Electrotechnical Committee.

[25] European Computer Manufacture Association <http://www.ecma.ch/>

[26] Comité Consultatif International Télégraphique & Téléphonique <http://www.sandybay.com/pc-web/CCITT.htm>

[27] Institute of Electrical & Electronic Engineers http://www.ieee.org/i3e_txhp.html

Reproduit dans une table de caractères codés, le codage à huit bits prendrait la forme d'une table à 7 bits miroitée à droite avec les mêmes emplacements pour les caractères de contrôle et les caractères graphiques (Tab.11). Parmi les différences entre les deux fausses moitiés de la table étendue, nous soulevons les faits suivants :

- ❑ Conformément à la logique de la pondération, la valeur « 1 » est attribuée au bit le plus important, le « B8 » qui crée la modification par rapport aux codes à 7 bits ;
- ❑ La notation x/y de la position des combinaisons binaires devient xx/yy ;
- ❑ Dans la moitié de droite, les emplacements des caractères ESPACE et DELETE (10/00 et 15/15) sont substitués par d'autres caractères graphiques si l'invocation est celle d'un jeu de 96 caractères ou laissés inutilisés dans le cas d'une invocation d'un jeu de 94 caractères seulement ;
- ❑ Quand désigné, ce qui n'est pas obligatoire dans tous les cas, un jeu de caractères de contrôles supplémentaires (CD) ne doit pas contenir le caractère d'échappement « ESCAPE » positionné au 01/11 ;
- ❑ Un jeu de caractères de contrôle supplémentaires (CD) ne doit pas inclure les caractères de contrôle de transmission spécifiés par ISO/IEC 6429 ;
- ❑ Contrôle de transmission spécifié par ISO/IEC 6429.

				B8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
				B7	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	
				B6	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	
				B5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	
B4	b3	B2	B1		00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
0	0	0	0	00	NUL	DLE	SP	0	@	P	^	P								
0	0	0	1	01	SUH	DC1	!	1	A	Q	A	Q								
0	0	1	0	02	STX	DC2	"	2	B	R	B	R								
0	0	1	1	03	ETX	DC3	#	3	C	S	C	S								
0	1	0	0	04	EOT	DC4	&	4	D	T	D	T								
0	1	0	1	05	ENQ	NAK	%	5	E	U	E	U								
0	1	1	0	06	ACK	SYN	&	6	F	V	F	V								
0	1	1	1	07	BEL	ETB	'	7	G	W	G	W								
1	0	0	0	08	BS	CAN	)	8	H	X	H	X								
1	0	0	1	09	HT	EM	(	9	I	Y	i	Y								
1	0	1	0	10	LF	SUB	*	:	J	Z	J	Z								
1	0	1	1	11	VT	ESC	+	;	K	[	K	{								
1	1	0	0	12	FF	ES4	,		L	\	L									
1	1	0	1	13	CR	ES3	-	=	M	]	m	}								
1	1	1	0	14	SO	ES2	.		N	^	n	~								
1	1	1	1	15	SI	ES1	/	?	O	_	o	DE L								

Tableau 12 : La structure étendue d'une table à 8 bits (ASCII étendu)

Un codage à 8 bits aurait ainsi la possibilité d'inclure par alternance dans la zone GD (10/01 - 15/14) des caractères accentués ou tout autre jeu de caractères d'une autre langue comme le grec ou l'arabe (le cas de l'ASMO 708). La technique de l'invocation (Locking shift) permettrait de faire entrer en vigueur un autre jeu de caractères à 8 bits, donc deux autres jeux de caractères à 7 bits permettant de la sorte de disposer en simultané de 4 jeux de caractères différents. Les 4 jeux peuvent être invoqués dans n'importe quelle partie (GG, GD) de la table des codes. La limite à 4

jeux de caractères simultanés est en fait arbitraire car d'autres techniques d'invocation peuvent être définies pour choisir un ou plusieurs jeux de caractères à n'importe quel moment du processus de traitement ou de transmission. Quatre jeux sont pourtant généralement suffisants pour répondre à toutes les situations d'un traitement multilingue. La problématique suivante à résoudre était celle de permettre une identification d'un jeu quelconque de caractères à n'importe quel moment de l'opération de traitement ou de transmission de manière automatique et d'en informer le récepteur du message pour identifier et reconnaître le jeu en question. Cette problématique sera largement traitée dans les nouveaux protocoles Internet sous les chapitres *étiquetage* et *négociation* de la langue des documents en hypertexte. Nous l'aborderons au niveau du point 4 de ce chapitre.

Pour faciliter le choix de l'un parmi les jeux des caractères disponibles, l'ISO a procédé à leur regroupement dans des registres internationaux identifiés sous la référence ISO-IR xx où xx est un numéro de série^[30]. L'IRV de l'ISO 646(1983) est identifié dans ce registre par le code ISO-IR 2 ; celui de l'ISO/IEC 646(1991) par le code ISO-IR 6 (Tab.8). Des codes supplémentaires peuvent être désignés dans ces registres pour être invoqués conformément à un besoin particulier.

Les techniques d'extension de codes fonctionnent par des combinaisons plus complexes comme les séquences d'échappement, les commandes de contrôle, les annonceurs (announcers), le maillage (nesting) des jeux des caractères codés. Nous nous contenterons de présenter le mécanisme de fonctionnement de la norme ISO/IEC 2022^[31] pour expliquer davantage les mécanismes d'invocation et de désignation de jeux des caractères multilingues et leurs fonctionnements respectifs. Cette technique a en fait toujours caractérisé l'environnement de la gestion multilingue des opérations de traitement et d'échange de l'information. L'expérience démontrera ensuite l'importance de ces techniques d'extension pour le passage aux mécanismes de codage sur 16 et 32 bits (2 et plusieurs octets) soit au niveau des normes nationales comme c'est le cas des normes ASMO 662 et 708 etc. ou au niveau des normes internationales comme c'est le cas des normes ISO 8859-x, Unicode et ISO/IEC 10646. La situation actuelle, comme il en sera fait détail dans les chapitres suivants, progresse plus rapidement sur la voie de la multiplicité linguistique universelle tant dans les univers du monoposte que ceux des réseaux d'information du type Internet et Intranet à base d'une normalisation universelle multilingue.

2.2 ISO 2022 : UNE NORME D'EXTENSION DE CODES

Pour aller au-delà des 128 uniques possibilités de l'ASCII ou de ses variantes nationales, des normes ISO ont été mises en place pour assurer l'alternance entre deux jeux de caractères ou plus. Un projet d'extension de codes a été imaginé puis normalisé sous la référence ISO/IEC 2022. L'idée était de permettre le codage de différents caractères par la même combinaison binaire.

2.2.1 Structure de la norme ISO 2022

Dans sa structure de base, une table des codes normalisée ISO/IEC 2022^[32] est conforme au modèle générique de toute table de code constituée de zones de caractères graphiques et de

[30] Le Registre International, qui contient actuellement plus de 180 jeux de caractères codés, est régie par la norme ISO/IEC 2357 et publié par l'ECMA qui fonctionne sous l'autorité de l'ISO.

[31] International Standard ISO/IEC 2022 : Information technology-Character code structure and extension techniques. 4è éd.. 1994.47 p.

[32] Character Sets - ISO/IEC 2022. EWOS Guide to Open Systems Specifications. <http://www.ewos.be/tg-cs/gis2022.htm>

caractères de contrôle. Des caractères de contrôle et des séquences d'échappement permettront de réaliser certaines fonctions de contrôle nécessaires pour l'invocation et la désignation des caractères ou des jeux des caractères particuliers.

ISO/IEC 2022 construit un code à partir d'une sélection des éléments de codes suivants :

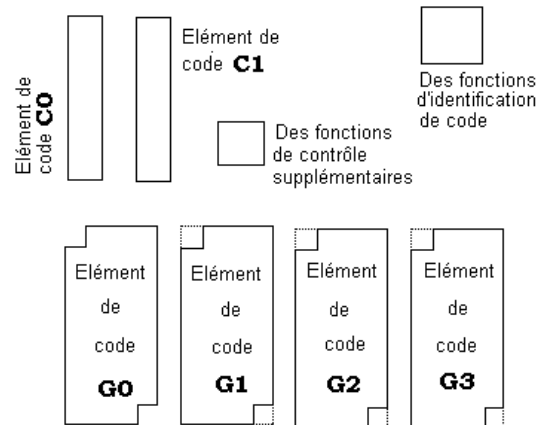


Figure 17 : Les éléments de code d'un jeu de caractères

• Des caractères graphiques

Quatre éléments de codes nommés G0, G1, G2 et G3 renferment les caractères graphiques nécessaires pour les zones GG et GD d'une table de codes à 8 bits ;

Les éléments de codes G1, G2 et G3 peuvent fournir chacun des attributions de codes à 94 ou 96 positions de caractères graphiques (Tab.13). Un jeu de 94 positions pourrait de la sorte attribuer des codes pour les positions 2/1 à 7/14 de la zone GG ou 10/1 à 15/14 de la zone GD, tout en excluant les positions allouées aux caractères ESPACE (2/0) et DEL (7/15) dans la zone GG et les deux positions miroirs correspondantes dans la zone GD (10/0 et 15/15 respectivement).

Un jeu de 96 positions permettrait par contre d'allouer des codes pour toutes les 96 positions des deux zones GG ou GD. Les éléments de codes G0 sont soumis aux mêmes conditions sauf qu'ils n'acceptent que 94 positions uniquement.

	0	1
0	NUL	DLE
1	SOH	DC1
2	STX	DC2
3	ETX	DC3
4	EOT	DC4
5	ENQ	NAK
6	ACK	SYN
7	BEL	ETB
8	BS	CAN
9	HT	EM
10	LF	SUB
11	VT	ESC
12	FF	IS4
13	CR	IS3
14	SO	IS2
15	SI	IS1

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
!	1	A	O	a	q											
"	2	B	H	b	r											
#	3	C	S	c	s											
\$	4	D	T	d	t											
%	5	E	U	e	u											
&	6	F	V	f	v											
'	7	G	W	g	w											
{	8	H	X	h	x											
}	9	I	Y	i	y											
*	:	J	Z	j	z											
+	;	K	[	k	{											
,	<	L	\	l												
-	=	M	]	m	}											
.	>	N	^	n	~											
/	?	O	_	o												

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
!	1	A	O	a	q											
"	2	B	H	b	r											
#	3	C	S	c	s											
\$	4	D	T	d	t											
%	5	E	U	e	u											
&	6	F	V	f	v											
'	7	G	W	g	w											
{	8	H	X	h	x											
}	9	I	Y	i	y											
*	:	J	Z	j	z											
+	;	K	[	k	{											
,	<	L	\	l												
-	=	M	]	m	}											
.	>	N	^	n	~											
/	?	O	_	o												

Tableau 13 : Les 32, 94 et 96 positions utilisées par le jeu C0 et les 4 jeux G0/G3

La démarche d'utilisation des jeux des caractères ainsi décrite est appelée selon les termes de l'ISO/IEC 2022 une « invocation ». Deux éléments de codes au plus parmi quatre (G0 à G3) peuvent être invoqués au même temps dans une table de codes à 8 bits chacun dans une des deux zones GG et GR. Par contre, un seul élément au plus est invoqué dans une table des codes à 7 bits qui d'ailleurs ne dispose pas de zone GD.

- **Les caractères de contrôle**

Deux éléments de codes nommés C0 et C1 contiennent des caractères de contrôle pour les besoins des zones CG et CD d'une table des codes à 8 bits ;

Les caractères de contrôle ont un nom et un acronyme d'identification sans pourtant disposer d'une représentation graphique. Les caractères BACKSPACE (BS=0/8), BELL (BEL=0/7), START OF HEADING (SOH=0/1) et ESCAPE (ESC=1/11) en sont des exemples. Ce sont des cas particuliers d'un concept plus générique : les fonctions de contrôle (Voir [Annexe B](#)).

Les éléments de codes C0 et C1 définissent les désignations des caractères de contrôle pour les 32 positions des caractères des zones CG et CD.

Il est nécessaire que l'élément de code C0 dispose du caractère ESCAPE (code d'échappement) dans la position 01/11 de sorte que son invocation n'annule pas sa présence comme caractère de contrôle.

Un jeu de caractères codés sur 8-bits qui dispose d'un élément de code C1 devrait normalement être invoqué dans la zone CD. Les caractères qui le constituent sont introduits par des séquences d'échappement et ne doivent donc pas inclure le caractère « ESCAPE ».

- **Les séquences d'échappement**

Dans ce cas, les séquences d'échappement fournissent les représentations codées des fonctions de contrôle supplémentaires définies dans la zone droite CD.

Il s'agit plus concrètement de deux caractères ou plus dont le premier est le caractère « ESCAPE » situé au 03/11 suivi soit d'un caractère intermédiaire et final ou d'un caractère final uniquement. Ces deux caractères définissent la fonction représentée par une séquence d'échappement quelconque. Les caractères intermédiaires sont choisis parmi les 16 emplacements de la colonne 02 ; les caractères finals appartiennent aux 79 positions restantes entre 03/00 et 07/14.

Les séquences d'échappement servent entre autres à désigner et invoquer des jeux de caractères graphiques et de contrôle dans les zones de la table de codes active.

2.2.2 Techniques d'invocation et de désignation de jeux de caractères à 7 et à 8 bits selon ISO/EC 2022

Pour indiquer lequel des caractères est visé par un code quelconque, on fait appel à une fonction de contrôle SHIFT ou une séquence d'échappement avec des effets analogues. Au moment de la lecture (ou de la réception), chaque fois qu'une fonction SHIFT (plusieurs sont définies) ou une séquence d'échappement est détectée, le contexte du lecteur change, et une table de code

différente est activée (Fig.18). Les tables de codes possibles peuvent être une version nationale de l'ISO 646 telle la table ASMO^[33] 449 normalisée ISO 9036. (Alt + shift Droit pour invoquer la table des codes arabes).

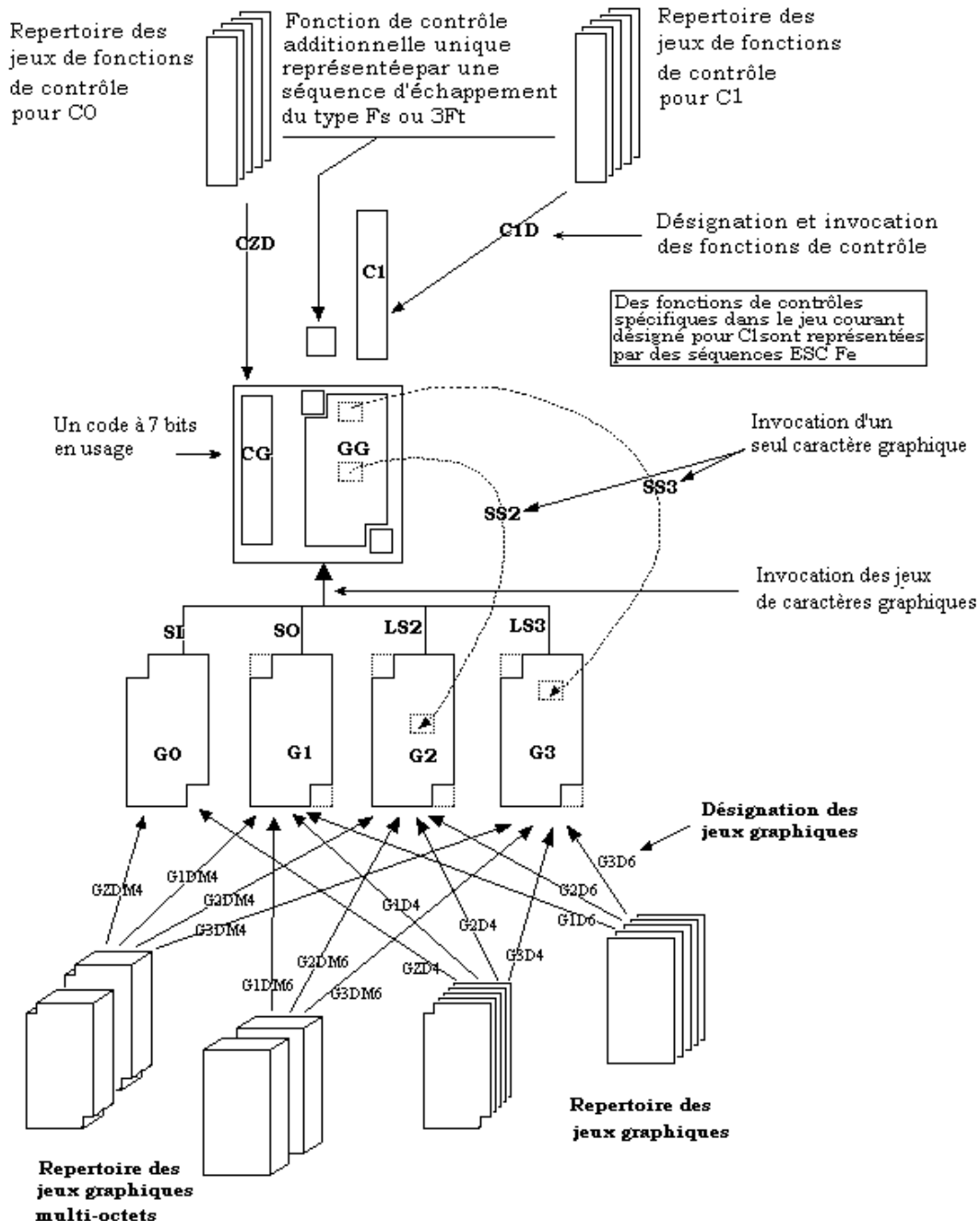


Figure 18 : Extension de code dans un environnement 7 bits

Source : ISO/IEC 2022:1994 Technologies de l'information -- Structure de code de caractères et techniques d'extension (Publiée actuellement en anglais seulement)

L'ISO/IEC 2022 fournit également les moyens pour coder un nombre presque illimité de caractères par une seule, mais pas unique, combinaison binaire. La norme ISO/IEC 2022 n'est pas réduite uniquement aux seules combinaisons à 7 bits. Elle a été étendue pour inclure des caractères codés sur 8 bits aussitôt que la structure de ceux-ci ait été définie dans la norme ISO 4873 (Fig.19).

[33] Arab Standards and Metrology Organization.

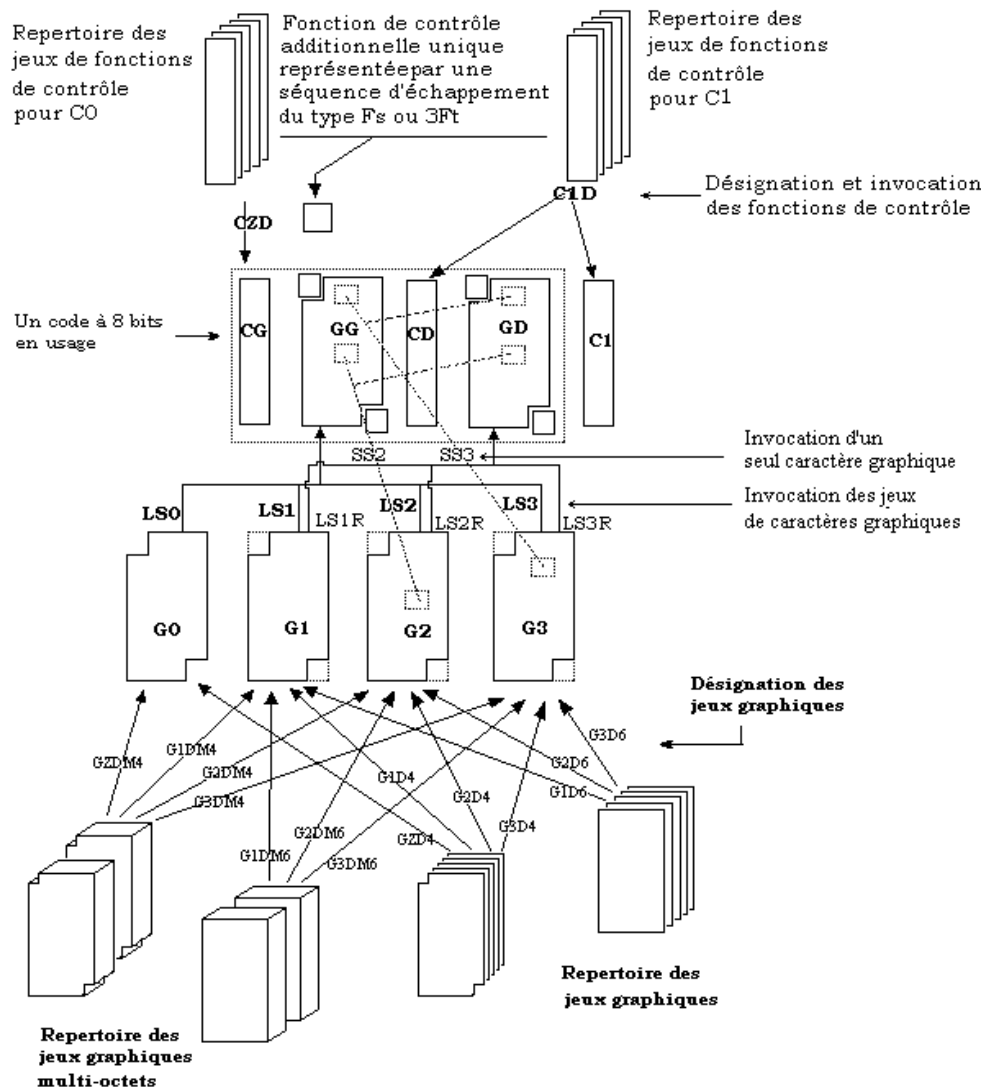


Figure 19 : Extension de code dans un environnement 8 bits

Source : ISO/IEC 2022:1994 Technologies de l'information -- Structure de code de caractères et techniques d'extension (Publiée actuellement en anglais seulement)

La norme ISO 4873 spécifie la structure des jeux des caractères codés sur 8-bits, sans toutefois définir une seule table de codes. Cette norme fixe le contenu de certaines zones, mais pour le reste des zones non affectées, seulement des options sont proposées. Pour le but de l'ISO/IEC 2022, les jeux des caractères sont identifiés par les désignations C0, C1, G0, G1. Les caractères de Contrôle occupent les colonnes 0,1 (C0), 8,9 (C1). Les colonnes 2-7 (G0) sont identiques à celle de l'ISO/IEC 646 (options incluses). Pour les colonnes 10-15 (G1), des options pour 94 ou 96 caractères sont spécifiées. Donc ISO 4873 est seulement le standard d'un standard générique qui prévoit 188 ou 190 caractères graphiques.

2.3 CONCLUSION

Le problème de l'ISO/IEC 2022 est que la lecture des données codées d'après cette norme exige des états de configuration machine limités pour des états de combinaisons binaires multiples. Par conséquent, son usage pratique n'a jamais été étendu. Avec l'avènement de l'infrastructure orientée 16 et 32 bits, des solutions partielles pour des problèmes urgents sont devenues

réalisables. Néanmoins, l'ISO/IEC 2022 fournit toujours la méthode de base où le changement des tables des codes est appliqué.

Dans le monde arabe des efforts similaires ont été entrepris pour se doter des artefacts nécessaires capables d'aligner la langue arabe à ce courant de normalisation dans le traitement et la transmission des données à échelles régionale et mondiale. La norme ASMO 449 ([Tab.18](#)) en est la concrétisation initiale suivie d'une série de normes (Voir [Annexe E](#)) gérant les extensions de codes des caractères arabes conformément aux évolutions enregistrées dans ce domaine à échelle internationale.

Notre étude portera à partir de ce point sur les particularités de cet aspect multilingue arabe/latin. Nous prendrons les accomplissements internationaux en matière d'extension de codes comme points de repères pour évaluer les initiatives nationales et régionales en terme d'arabisation des systèmes d'information et de communication de l'information scientifique et technique multilingue. Des points forts ainsi que des points faibles au niveau des techniques et surtout des usages (nous en aborderons les aspects ergonomiques des interfaces Homme-Machine) marqueront le parcours de cette expérience arabe dans l'instauration d'un contexte normatif durable mais encore critiquable.

3. CODAGE, NORMALISATION ET MULTILINGUISME ARABE/LATIN

Dans les pays de langue arabe, l'arabisation de l'informatique, bien que ses premiers balbutiements datent depuis les années 60 avec les systèmes à cartes perforées, n'a pu établir des standards de fait que depuis la fin des années 80. C'était suite à des initiatives très disparates dans le temps et dans l'espace étant donnée que les jeux des caractères utilisés étaient divers non seulement d'un pays à un autre, mais aussi d'un secteur à un autre. C'était généralement des jeux qui omettaient la voyellation et ne permettaient pas la cohabitation avec le caractère latin. D'ailleurs, le problème de l'articulation des mots arabes et leur voyellation restent encore sans solutions adéquates. Bien qu'elles semblent être une problématique d'importance seconde par comparaison à la mise en place d'une norme internationale solide pour la représentation du caractère, l'articulation et la voyellation de certains termes arabes constituent encore un problème qui mérite un intérêt particulier.

3.1 L'ASMO 449 : UN DEBUT DE NORMALISATION

Depuis la deuxième moitié des années 70, il y avait eu un début de consensus arabe pour harmoniser les tentatives de normalisation entreprises de façon disparate. C'était en 1975 que les spécialistes en la matière se sont réunis pour la première fois (à l'I.E.R.A. au Maroc) pour débattre de la possibilité d'établir un code standard unifié pour la transmission des données en langue arabe qui respecterait les caractéristiques suivantes :

- Établir un alphabet fonctionnel avec un minimum de signes utilitaires;
- Établir un alphabet complet avec les signes de voyellation;
- Établir un alphabet compatible avec l'alphabet international N°5 respectant les spécificités de la norme ISO/IEC 646.

Ce n'est qu'en 1976 que le premier CODE ARabe de transmission des données, CODAR 1 ([Tab.14](#)), a vu le jour comme contrepartie arabe du code ASCII. Il a été immédiatement soumis à l'examen auprès de toutes les parties intéressées qui se sont rassemblées durant la Réunion

Régionale pour les Pays Arabes sur la Stratégie et les Politiques Informatiques organisée à Bizerte (Tunisie) en Juin 1996. CODAR 1 a été alors confronté à un autre projet, en l'occurrence un code irakien parallèle.

Codar 1					b7	0	0	0	0	1	1	1	1
					b6	0	0	1	1	0	0	1	1
					b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1			0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	ء	ط	ء	ء	ء
0	0	0	1	1	SOH	DC1	!	1		ظ	ء	ء	ء
0	0	1	0	2	STX	DC2	"	2	ب	ء	ء	ء	ء
0	0	1	1	3	ETX	DC3	#	3	ت	ء	ء	ء	ء
0	1	0	0	4	EOT	DC4	\$	4	ث	ف	ا	ء	ء
0	1	0	1	5	ENQ	NAK	%	5	ج	ق	ت	ء	ء
0	1	1	0	6	ACK	SYN	&	6	ح	ك	ء	ء	ء
0	1	1	1	7	BEL	ETB	'	7	خ	ل	ء	ء	ء
0	0	0	0	8	BS	CAN	(	8	د	م	ء	ء	ء
1	0	0	1	9	HT	EM	)	9	ذ	ن	ة	ء	ء
1	0	1	0	10	LF	SUB	*	:	ر	ه	و	ء	ء
1	0	1	1	11	VT	ESC	+	:	ز	و	ى	ء	ء
1	1	0	0	12	FF	FS	,	<	س	پ	ي	ء	ء
1	1	0	1	13	CR	GS	—	=	ش	ء	پ	ء	ء
1	1	1	0	14	SO	RS	.	>	ص	ء	ء	ء	ء
1	1	1	1	15	SI	US	/	؟	ض	ء	ظ	ء	ء

Tableau 14 : CODAR-1

Les confrontations portaient sur la nécessité ou la non nécessité de la voyellation dans le jeu alphabétique arabe. Quelques réarrangements ont été apportés sur CODAR 1 pour aboutir en fin de compte à un projet provisoire : CODAR 2 (Tab.15).

Codar 2					b7	0	0	0	0	1	1	1	1
					b6	0	0	1	1	0	0	1	1
					b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1			0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	ء	ء	ب	ظ
0	0	0	1	1	SOH	DC1	!	1	ء	ء	ء	ت	ء
0	0	1	0	2	STX	DC2	"	2	ء	ء	ء	ة	ء
0	0	1	1	3	ETX	DC3	#	3	ء	ء	ء	ث	ف
0	1	0	0	4	EOT	DC4	\$	4	ء	ء	ء	ج	ق
0	1	0	1	5	ENQ	NAK	%	5	ء	ء	ء	ح	ك
0	1	1	0	6	ACK	SYN	&	6	ء	ء	ء	خ	ل
0	1	1	1	7	BEL	ETB	'	7	ء	ء	ء	د	م
0	0	0	0	8	BS	CAN	(	8	ء	ء	ء	ذ	ن
1	0	0	1	9	HT	EM	)	9	ء	ء	ء	ر	ه
1	0	1	0	10	LF	SUB	*	:	ء	ء	ء	ز	و
1	0	1	1	11	VT	ESC	+	:	ء	ء	ء	س	پ
1	1	0	0	12	FF	FS	,	<	ء	ء	ء	ش	ء
1	1	0	1	13	CR	GS	—	=	ء	ء	ء	ص	ي
1	1	1	0	14	SO	RS	.	>	ء	ء	ء	ض	ء
1	1	1	1	15	SI	US	/	؟	ء	ء	ء	ظ	DEL

Tableau 15 : CODAR-2

Ce projet a été ensuite soumis respectivement à la Conférence CASTARAB (Conférence des Ministres chargés de l'application de la Science et de la Technologie au Développement) tenue à Rabat (Maroc) en août 1976, et à la 19^e session de la Conférence Générale de l'UNESCO tenue en décembre 1976 à Nairobi qui l'ont entièrement adopté (Résolution 222 de l'UNESCO).

En juin 1977, le COARIN, comité créé par l'IBI (Bureau Intergouvernemental pour l'Informatique) pour mettre en place une norme commune d'utilisation de la langue arabe en informatique, organise une réunion à Rome durant laquelle les controverses relatives aux différents aspects de la codification du jeu des caractères arabes aboutissent à un code unifié. Le CODAR-U a été adopté le 15 juin 1977 et ensuite normalisé sous la référence ISO N°59 le 1er juin 1982 (Tab.16).

CODAR-U					b7	0	0	0	0	1	1	1	1
					b6	0	0	1	1	0	0	1	1
					b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1			0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	°	ب	ظ	
0	0	0	1	1	SOH	DC1	!	1	⎵	°	ت	ع	
0	0	1	0	2	STX	DC2	"	2	⎵	°	ة	غ	
0	0	1	1	3	ETX	DC3	#	3	⎵	°	ث	ف	
0	1	0	0	4	EOT	DC4	\$	4	⎵	°	ج	ق	
0	1	0	1	5	ENQ	NAK	%	5	⎵	°	ح	ك	
0	1	1	0	6	ACK	SYN	&	6	⎵	°	خ	ل	
0	1	1	1	7	BEL	ETB	'	7	⎵	°	د	م	
0	0	0	0	8	BS	CAN	(	8	⎵	°	ذ	ن	
1	0	0	1	9	HT	EM	)	9	⎵	°	ر	هـ	
1	0	1	0	10	LF	SUB	*	:	⎵	°	ز	و	
1	0	1	1	11	VT	ESC	+	:	⎵	°	س	ي	
1	1	0	0	12	FF	FS	،	<	⎵	°	ش	ا	
1	1	0	1	13	CR	GS	—	=	⎵	°	ص	ى	
1	1	1	0	14	SO	RS	.	>	⎵	°	ض	ق	
1	1	1	1	15	SI	US	/	?	⎵	°	ط	DEL	

Tableau 16 : CODAR-U

Entre-temps, le CODAR-U a été présenté par plusieurs instances arabes et internationales dont les plus influentes : l'Union Arabe des Télécommunications (UAT) et l'Organisation Arabe de Normalisation et de Métrologie (ASMO). Celle-ci se saisit du dossier en 1980 pour lequel elle a créé un Comité Technique (Comité N°8) qui aura ensuite un impact très direct sur la genèse d'une série de normes subsidiaires.

Devant ces agitations normatives qui ont perturbé l'activité commerciale des constructeurs de matériel informatique destiné aux pays arabes, le *European Computer Manufacturers Association* (ECMA), autorité d'enregistrement de l'ISO pour les codes informatiques, a proposé en février 1982 un code pléthorique destiné à satisfaire les attentes des partenaires arabes. Tous les protagonistes arabes, convoqués par l'ASMO et l'ALECSO se sont réunis à l'IERA, Rabat

(Maroc) les 22 et 23 avril 1982 pour se mettre d'accord sur un code unifié et définitif, en l'occurrence le CODAR-UF/FD (Code arabe unifié, Forme Définitive) ([Tab.17](#)).

Codar-U/FD					b ₇ 0 0 0 0 1 1 1 1								
					b ₆ 0 0 1 1 0 0 1 1								
					b ₅ 0 1 0 1 0 1 0 1								
					0 1 2 3 4 5 6 7								
b ₄	b ₃	b ₂	b ₁										
0	0	0	0	0	NUL	DLE	SP	0	@	ذ	—	/	
0	0	0	1	1	SOH	DC1	!	1	ء	ر	ف	س	
0	0	1	0	2	STX	DC2	"	2	آ	ز	ق	ه	
0	0	1	1	3	ETX	DC3	#	3	ا	س	ك		
0	1	0	0	4	EOT	DC4	␣	4	ؤ	ش	ل		
0	1	0	1	5	ENQ	NAK	%	5	!	ص	م		
0	1	1	0	6	ACK	SYN	Ⓢ	6	ث	ض	ن		
0	1	1	1	7	BEL	ETB	/	7	ا	ط	هـ	÷	
1	0	0	0	8	BS	CAN	)	8	ب	ظ	و	×	
1	0	0	1	9	HT	EM	(	9	ة	ع	ى	«	
1	0	1	0	10	LF	SUB	*	:	ت	غ	پ	»	
1	0	1	1	11	VT	ESC	+	؛	ث	]	/	}	
1	1	0	0	12	FF	IS4	Ⓢ	>	ج	\	فـ		
1	1	0	1	13	CR	IS3	_	=	ر	[	مـ	{	
1	1	1	0	14	SO	IS2	.	<	خ	^	/	-	
1	1	1	1	15	SI	IS1	/	؟	ن	—	و		

Tableau 17 : La version définitive Codar-U/FD

Avant d'être adoptée définitivement comme norme ASMO 449, cette version définitive a été rectifiée au niveau des positions suivantes (Voit [Tab.18](#)) :

- ❑ Les caractères 4/0 (@), 5/14 (^) et 6/0 (␣) ont gardé leurs formes latines afin d'être utilisés pour des fonctions de contrôle ;
- ❑ Le caractère & est gardé avec les mêmes fonctions qui lui sont attribuées par le contexte latin et remplacera la virgule arabe à la position 2/6 ;
- ❑ La virgule arabe (ex-position 2/6) est située à la position 2/12 ;
- ❑ Un caractère de ligature, indispensable pour le style de l'écriture arabe cursive, est défini à la position 6/0 ;
- ❑ Le caractère 7/7 (division) est remplacé par le caractère 2/15 (barre oblique) ;

- ❑ Le caractère 7/8 (multiplication) est remplacé par le caractère 10/2 (Astérisque) ;
- ❑ Les caractères 7/9 et 7/10 (guillemets français d'ouverture et de fermeture) sont remplacés par le caractère 2/2 (guillemets anglais).

				b ₇	0	0	0	0	1	1	1	1
				b ₆	0	0	1	1	0	0	1	1
				b ₅	0	1	0	1	0	1	0	1
					0	1	2	3	4	5	6	7
b ₄	b ₃	b ₂	b ₁		0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	ذ	—	/
0	0	0	1	1	SOH	DC1	!	1	ء	ر	ف	س
0	0	1	0	2	STX	DC2	"	2	آ	ز	ق	ه
0	0	1	1	3	ETX	DC3	#	3	ا	س	ك	
0	1	0	0	4	EOT	DC4	¤	4	ؤ	ش	ل	
0	1	0	1	5	ENQ	NAK	%	5	!	ص	م	
0	1	1	0	6	ACK	SYN	&	6	ئ	ض	ن	
0	1	1	1	7	BEL	ETB	/	7	ا	ط	هـ	
1	0	0	0	8	BS	CAN	)	8	ب	ظ	و	
1	0	0	1	9	HT	EM	(	9	ة	ع	ى	
1	0	1	0	10	LF	SUB	*	:	ت	غ	ي	
1	0	1	1	11	VT	ESC	+	؛	ث	]	//	}
1	1	0	0	12	FF	IS4	'	>	ج	\	ف	ا
1	1	0	1	13	CR	IS3	_	=	ح	[	=	{
1	1	1	0	14	SO	IS2	.	<	خ	^	/	-
1	1	1	1	15	SI	IS1	/	؟	د	—	و	DEL

Tableau 18 : Le jeu de caractères arabe (ASMO 449)

C'était le départ d'une série de normes subsidiaires élaborées dans le cadre du Comité N°8 de l'ASMO qui a mis au point une série de codes appropriés à la langue arabe (Voir [Annexe E](#)).

3.2 ASMO 662 = ASCII ETENDU (LATIN 2)

L'enthousiasme pour la conception d'une norme qui correspondrait intégralement au standard international ISO/IEC 646, et qui permettrait par conséquent de préparer la mise en place d'un système arabe indépendant, a abouti en 1985 à la définition d'une nouvelle norme étendue de codage du caractère arabe sur 8 bits ([Tab.19](#)) destinée à servir de plate-forme d'échange des données exclusivement entre des systèmes informatiques de traitement des données en langue arabe. Cette initiative visait les objectifs suivants :

- ❑ La mise en place d'une table de codes arabes indépendante qui anticiperait la conception d'un ordinateur unilingue arabe ;
- ❑ La possibilité d'étendre le jeu des caractères arabes en fonction des exigences nationales et internationales sans besoin d'une désignation anticipée stricte. Ceci est d'ordre à permettre aux langues utilisant le caractère arabe d'y retrouver la solution adéquate à l'automatisation de leurs langues respectives (i.e. le persan, l'afghan, le berbère...) ;
- ❑ Éliminer la redondance qui pourrait résulter de l'usage d'une table commune pour les caractères arabes et latins ;
- ❑ Éviter le recours aux programmes de l'analyse contextuelle pour gérer la directionnalité de certaines paires de signes communes aux environnements arabe et latin (Tab.22).

La norme ASMO 662^[34] reprend intégralement le jeu de caractères codés selon la norme ASMO 449 auquel elle applique les techniques d'extension des codes conformément à la norme ISO/IEC 2022 et la norme ISO/IEC 4873.

				0	1	0	0	0	0	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Tableau 19 : La norme ASMO 662

Pour des considérations géopolitiques et économiques, il n'a pas été possible de mettre en place une politique d'industrialisation d'un ordinateur arabe unilingue. Le contexte d'usage multilingue de l'informatique et l'effet de la mondialisation vont désormais à l'encontre d'une pareille tendance. Dans un moment où les systèmes distribués d'information et les réseaux téléinformatiques prennent une dimension universelle guidée par une primauté et une lourdeur stratégique et économique occidentale et plus particulièrement anglo-saxonne, la contre partie latine était devenue un passage obligé pour permettre aux autres langues de marquer leur présence dans le contexte international. Le penchant était alors orienté vers l'adoption d'une

[34] ASMO. Traitement de l'information - Jeux de caractères arabes codés à 8 éléments pour l'échange de l'information : Norme arabe n° 662 - 1985. Ligue des états arabes. 9p.

norme de codage étendu à 8 bits pour deux jeux de caractères arabe et latin : la norme ASMO 708 est née.

3.3 ASMO 708 (ISO/IEC 8859-6 = ECMA 114)

Conçue pour gérer un bilinguisme arabe latin, la norme ASMO 708 attribue à la zone GG un jeu de caractères latins conforme à la norme ISO/IEC 646 et à la zone GD un jeu de caractères définis dans la norme ASMO 449 avec deux particularités essentielles (Tab.20) :

- ❑ Définir un seul emplacement pour les caractères identiques dans les deux normes en question. Le résultat, présenté dans le tableau suivant, est le côtoïement d'un jeu GD de 50 caractères graphiques arabes (ASMO 449 moins les caractères communs) avec un jeu GG de 128 caractères latins ;
- ❑ Appliquer les techniques d'extension de codes conformément à la norme ISO/IEC 2022 et ISO/IEC 4873.

ASMO 708

00000000	00000001	00000010	00000011	00000100	00000101	00000110	00000111	00001000	00001001	00001010	00001011	00001100	00001101	00001110	00001111	00010000	00010001	00010010	00010011	00010100	00010101	00010110	00010111	00011000	00011001	00011010	00011011	00011100	00011101	00011110	00011111	00100000	00100001	00100010	00100011	00100100	00100101	00100110	00100111	00101000	00101001	00101010	00101011	00101100	00101101	00101110	00101111	00110000	00110001	00110010	00110011	00110100	00110101	00110110	00110111	00111000	00111001	00111010	00111011	00111100	00111101	00111110	00111111	01000000	01000001	01000010	01000011	01000100	01000101	01000110	01000111	01001000	01001001	01001010	01001011	01001100	01001101	01001110	01001111	01010000	01010001	01010010	01010011	01010100	01010101	01010110	01010111	01011000	01011001	01011010	01011011	01011100	01011101	01011110	01011111	01100000	01100001	01100010	01100011	01100100	01100101	01100110	01100111	01101000	01101001	01101010	01101011	01101100	01101101	01101110	01101111	01110000	01110001	01110010	01110011	01110100	01110101	01110110	01110111	01111000	01111001	01111010	01111011	01111100	01111101	01111110	01111111	10000000	10000001	10000010	10000011	10000100	10000101	10000110	10000111	10001000	10001001	10001010	10001011	10001100	10001101	10001110	10001111	10010000	10010001	10010010	10010011	10010100	10010101	10010110	10010111	10011000	10011001	10011010	10011011	10011100	10011101	10011110	10011111	10100000	10100001	10100010	10100011	10100100	10100101	10100110	10100111	10101000	10101001	10101010	10101011	10101100	10101101	10101110	10101111	10110000	10110001	10110010	10110011	10110100	10110101	10110110	10110111	10111000	10111001	10111010	10111011	10111100	10111101	10111110	10111111	11000000	11000001	11000010	11000011	11000100	11000101	11000110	11000111	11001000	11001001	11001010	11001011	11001100	11001101	11001110	11001111	11010000	11010001	11010010	11010011	11010100	11010101	11010110	11010111	11011000	11011001	11011010	11011011	11011100	11011101	11011110	11011111	11100000	11100001	11100010	11100011	11100100	11100101	11100110	11100111	11101000	11101001	11101010	11101011	11101100	11101101	11101110	11101111	11110000	11110001	11110010	11110011	11110100	11110101	11110110	11110111	11111000	11111001	11111010	11111011	11111100	11111101	11111110	11111111

Tableau 20 : La norme ASMO 708 (ISO/IEC 8856-6 / ECMA 114)

Étant donnée la dimension combinatoire entre deux jeux de caractères très utilisés dans les applications scientifiques, cette norme a été adoptée par l'ECMA sous la référence ECMA 114. L'ISO l'a également adopté pour sa nouvelle série de normes étendues 8859 sous la référence ISO/IEC 8859-6.

Il convient à ce niveau de s'arrêter sur cette dernière norme ISO/IEC 8859 pour mettre en exergue le rôle important qu'elle a joué dans la mise en place d'une solution multilingue universelle.

3.3.1 ISO/IEC 8859-x

La norme ISO/IEC 8859 a été développée pour reproduire une seule et unique représentation binaire des caractères graphiques. Au fait, l'ISO 8859 n'est pas un standard, mais une famille entière de normes qui commencent par ce qui était appelé « ASCII à huit bits » communément connu par le terme « Alphabet Latint-1 » (ISO8859/1). Seulement, étant donné qu'il n'est pas facile de représenter tous les caractères désirés dans une table de 94+96 codes, l'ISO/IEC 8859 est subdivisée en plusieurs parties, chacune définie pour une région particulière du monde afin de servir les exigences et les besoins des groupes de langues (Europe: Ouest, Est, Nord, Vers le sud; Cyrillique, Grec, Arabe, Hébreu). Chaque partie contient les 94 (G0) de l'ASCII comme un sous-ensemble par défaut, complétée par un jeu de 96 (G1) variables^[35] (Tab.21). Donc le code de l'ISO/IEC 8859 est seulement unique dans un sens restreint. Car au moment où les graphiques de régions différentes seront combinés dans un texte, les techniques de l'ISO/IEC 2022 pour gérer le changement des tables des jeux de codes (mécanisme SO/SI de séquences d'échappement ISO/IEC 2022 en mode 7-bits) deviennent obligatoires.

	ı	ƒ	£	¥	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

**ISO/IEC 8859-1 : Europe de l'Ouest,
Amérique Latine**

	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
°	q	r	s	t	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
Ř	Š	Ť	Ž	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
Ø	Ň	Ů	Ű	Ų	Ŵ	Ŷ	Ÿ	ı	ı	ı	ı	ı	ı	ı	ı
ř	š	ť	ž	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
đ	ñ	ň	ő	û	ŵ	ŷ	ÿ	ı	ı	ı	ı	ı	ı	ı	ı

ISO/IEC 8859-2 : Europe de l'Est

	Ĥ	Ħ	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
°	h	e	z	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
Ā	Ă	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ
Ñ	Ō	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ	Ȯ
ā	ă	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ
ñ	ō	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ	ȯ

**ISO/IEC 8859-3 : Europe/divers (Esperanto,
Maltais etc.)**

	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
°	q	r	s	t	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
Ā	Ă	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ	Ȧ
Ø	Ň	Ů	Ű	Ų	Ŵ	Ŷ	Ÿ	ı	ı	ı	ı	ı	ı	ı	ı
ā	ă	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ	ȧ
đ	ñ	ň	ő	û	ŵ	ŷ	ÿ	ı	ı	ı	ı	ı	ı	ı	ı

**ISO/IEC 8859-4 : Pays Scandinaves &
Baltiques (couverts aussi par 8859-1)**

	Ё	Ѣ	Ѧ	Ѩ	Ѭ	Ѯ	Ѱ	Ѳ	Ѵ	Ѷ	Ѹ	Ѻ	Ѽ	Ѿ	ѿ
А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
Ŧ	ё	ѣ	ѧ	ѩ	ѭ	ѯ	ѱ	ѳ	ѵ	ѷ	ѹ	ѻ	ѽ	ѿ	ѿ

ISO/IEC 8859-5 : Cyrillique

ISO/IEC 8859-6 : Arabe

[35] IBM a suivi ce modèle en étendant son code EBCDIC pour contenir les mêmes 190 caractères graphiques.

	´	ˆ	£			ı	§	ˆ	ø		«	¬	-		-
°	±	²	³	´	ˆ	À	•	É	Η	Ί	»	Ό	×	Υ	Ω
Ά	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
Π	Ρ		Σ	Τ	Υ	Φ	Χ	Ψ	Ω	Ϊ	Ϋ	ά	έ	ή	ί
ύ	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο
π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	

ISO/IEC 8859-7 : Grec

		Φ	£	¥	ı	§	ˆ	ø	×	«	¬	-	ø	-
°	±	²	³	´	μ	¶	•	ˆ	÷	»	¼	½	¾	
														=
×	ı	λ	τ	π	ı	ı	π	ı	ı	ı	ı	ı	ı	ı
ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı

ISO/IEC 8859-8 : Hébreu

	ı	Φ	£	¥	ı	§	ˆ	ø	×	«	¬	-	ø	-
°	±	²	³	´	μ	¶	•	ˆ	÷	»	¼	½	¾	ı
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ

ISO/IEC 8859-9 : Latin 5 identique à 8859-1
sauf que le turc remplace l'islandais

	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
°	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ
Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ	Ǻ

ISO/IEC 8859-10 : Latin 6
(Lapons/Nordiques/Langues esquimaudes)

Tableau 21 : La série de normes ISO/IEC 8859-x

3.3.2 ISO/IEC 8859-6

Les particularités de l'ISO/IEC 8859-6 sont discutées à travers les points suivants :

La redistribution des codes :

- La position 02/04 a été attribuée au signe Dollar (\$) suite à l'affectation du signe monétaire international (؄) à la position 10/04 dans la zone GD pour désigner le symbole monétaire arabe concerné ;
- La position 07/14 a été attribuée au signe non espacé tilde (~) conformément à la norme ISO 6937-2 ;
- La position 10/00 est attribuée au caractère Espace insécable (NBSP=Non Break Space) ;
- La position 10/13 est attribuée au caractère du tiret temporaire de coupure de ligne (SHY).

Les caractéristiques d'usage :

- **Le problème des chiffres** : Il est possible d'utiliser les chiffres indiens positionnés en parallèle aux chiffres arabes dans les mêmes emplacements entre 03/00 et 03/09. Une grande polémique existe encore au niveau de l'appellation et de l'usage des chiffres arabes et hindous. À part l'usage non normalisé des mêmes chiffres par la communauté arabe, un écart de terminologie est également observé dans les documents officiels de la normalisation des caractères. Alors que le texte de la norme arabe ASMO 709 fait référence aux « caractères indiens et arabes » dans son alinéa 4.3.5^[36], la norme ISO8859-6 rapporte que « les caractères alloués aux positions 03/00 à 03/09 figurent

[36] ASMO. Jeux de caractères arabes/latins codés à 8 éléments pour l'échange de l'information : Normes arabe n° 708 - 1988. Ligue des états arabes. 6p.

sous forme de deux jeux différents de symboles graphiques. Le jeu de symboles graphiques de gauche est celui de l'écriture latine, celui de droite est celui de l'écriture arabe »^[37]. Nous reprendrons cet aspect de l'usage des chiffres dans un chapitre suivant.

- ❑ **Le problème de la bidirectionnalité** : Partant de l'opposition entre les langues latines et la langue arabe dans le sens de l'écriture (gauche-droite et droite-gauche respectivement), la norme ISO/IEC 8859-6 prévoit l'usage de fonctions de commandes spécifiques pour gérer le changement de la directionnalité de l'écriture ou de l'insertion des valeurs numériques. Ces deux aspects nécessiteront de notre part une analyse particulière pour étudier leur implication sur l'ergonomie et l'usage des interfaces Homme-Machine dans un contexte d'information et de communication multilingue.
- ❑ **Les caractères surimprimés ou composés (overstriking)** : Contrairement à l'ISO 646, l'ISO 8859 interdit en général le recours à des fonctions de commande, telles que Espace Arrière (Back Space) ou Retour Chariot (Carriage Return) pour la représentation codée des caractères composés. Cette technique constitue l'une des approches d'extension de l'ensemble du jeu des caractères, initialement prévue pour contourner la complexité de la méthode ISO/IEC 2022 dans la formation des caractères marqués par des symboles diacritiques. C'est le modèle utilisé par le CCITT pour le télex européen et par le format MARC et les systèmes des bibliothèques automatisées. Il est aussi utilisé efficacement par les systèmes des bibliothèques NOTIS et DOBIS. Il n'a jamais été accepté par les fabricants du matériel de traitement des données qui ont préféré une forme entièrement différente d'extension des caractères basée sur un seul emplacement par caractère (chaque caractère accentué occupe une place indépendante dans le code).

La normalisation arabe applique une technique différente. En l'absence de signes diacritiques (voyellation) comme symboles particuliers à un caractère particulier, la table des codes arabes traite les signes de voyellation (de 6/11 à 7/2) comme caractères indépendants. Ceci justifie le fait qu'à l'opposé de l'écriture latine où les signes non espacés (tréma, accents...) précèdent le caractère de base (espacé), l'écriture arabe procède à l'écriture du caractère graphique de base suivi de son articulation comme caractère graphique indépendant non espacé^[38].

- ❑ **La ligature** : Beaucoup de critiques ont été formulées à l'égard de l'absence de certains caractères comme le « œ » français. En fait, « œ » n'est pas un caractère mais plutôt une ligature (une combinaison de deux caractères à des fins d'impression et d'affichage). Les ligatures ne font pas partie de l'ISO 8859-X.
- ❑ **L'analyse contextuelle** : Cette technique permet d'identifier la forme et la direction de l'écriture arabe à deux niveaux. D'une part elle définit par des algorithmes d'analyse de contexte la forme des caractères à adopter en fonction de leurs emplacements dans le mots (début, milieu, fin, isolé). Elle permet également de spécifier l'alternance entre l'un et l'autre des paires des caractères symétriques dans les deux contextes linguistiques comme illustré dans le tableau suivant. Cette technique aura également des répercussions sur l'ergonomie de l'interface Homme-Machine multilingue comme nous le verrons dans le chapitre suivant.

[37] ISO 8859. Traitement de l'information - jeux de caractères graphiques codés sur un seul octet : partie 6, alphabet latin/arabe. 1987. 5p.

[38] Ahmed ABOULHIJA. Muwasafat wa makais litaarib al maaloumatia=Normes et standards pour arabiser l'informatique. In : Istikhdam Allougha Al Arabia fi al Maaloumatia = Usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.150-188.

Position	Glyphe	Equivalents ISO 646	Equivalents ASMO
2/8	Parenthèse ouverte	(	)
2/9	Parenthèse fermée	)	(
2/12	Virgule	,	٬
3/11	Point virgule	;	؛
3/12	Plus petit que	<	>
3/14	Plus grand que	>	<
3/15	Point d'interrogation	?	؟
5/11	Crochet ouvert	[	]
5/13	Crochet fermé	]	[
7/11	Accolade ouverte	{	}
7/13	Accolade fermée	}	{

Tableau 22 : Paires des caractères symétriques arabes/latins

3.4 LES JEUX DES CARACTERES CODES SUR PLUSIEURS OCTETS

Nous avons évoqué jusqu'à présent une série de normes et de standards de codification des caractères qui, en plus de la diversité dans leurs structures internes, présentent une différenciation supplémentaire au niveau de leur applicabilité linguistique et culturelle. Ces nuances sont exprimées à travers leurs particularités linguistiques en servant une seule langue : i.e. l'hébreu (ISO/IEC 8859-8) ou leurs spécificités scripturales en servant un groupe de langues : i.e. langues latines (ISO 646, ISO/IEC 8859-1...), l'arabe (les normes ASMO et équivalents ISO), le cyrillique (ISO/IEC 8859-5).

Une nouvelle catégorie est née : les normes à vocation universelle qui supportent la plupart des langues écrites du monde. Un tournant stratégique dans le traitement et surtout l'échange de l'information multilingue est en train de se concrétiser.

En effet, de nos jours, les jeux des caractères codés sur plusieurs octets attirent beaucoup l'attention. Selon Hudrisier, « *le besoin de coder les caractères sur plusieurs octets est d'abord né au Japon. Les « langues partiellement ou entièrement idéographiques » (chinois, japonais, coréen) ne pouvaient pas, à l'évidence, fonctionner, sur les solutions de bricolage informatique qui peuvent encore provisoirement satisfaire les langues transcrites en écritures alphabétiques et syllabaires. C'est ainsi que se sont mis en place, d'abord des solutions spécifiques comme le Japan Informatic Standard (JIS), puis un projet industriel de représentation universelle des caractères sur 16bits (standard industriel Unicode), qui est normalisé comme étant le premier plan à 2 octets (le BMP, Basic Multilingual Plane, Plan Basique Multilingue), d'une norme plus universelle (canoniquement à 4 octets), c'est la norme ISO-IEC/10646* »^[39].

Cependant, s'il peut paraître qu'il s'agit là d'une nouvelle forme de codage clairement définie, à notre sens, il ne s'agit pas d'un nouveau code du tout car comme nous l'avons déjà vu avec les techniques d'extensions des codes et des caractères composés, des schémas multi-octets ont déjà été proposés depuis longtemps :

- ❑ *Celui de l'ISO 646.* Certains caractères peuvent être représentés par 1,3,5 octets ou plus, par l'usage des séquences telles que : « char » BACKSPACE « char » ... ;
- ❑ *Celui de l'ISO 6937-2.* Les caractères graphiques peuvent être formés par des diacritiques en plus des lettres, donnant lieu à une représentation mixte d'octet simple/double ;

[39] Henri HUDRISIER. La norme de représentation des caractères sur plusieurs octets : (ISO-IEC/10646), un enjeu pour le Sommet d'Hanoi. MED'IACOM INFOROUTES : un défi pour le monde arabe, la Méditerranée et la Francophonie. Université de Paris VIII. Beyrouth les 9, 10 et 11 avril 1997

- ❑ Le schéma utilisé par les standards nationaux chinois, japonais et coréen. Avant l'unification « Han »^[40], les caractères dans ces langues asiatiques se basaient sur une représentation des caractères graphiques sur deux octets, chacun pris dans un jeu de caractères de 7 bits d'une matrice de 94 positions semblable à celle de l'ASCII.

En plus de ces schémas possibles de codification, plusieurs possibilités ont été inventées et utilisées au Japon et en Chine par un mixage des représentations des caractères simples et/ou doubles dans des variations très complexes. Cet aspect dépasse les limites de notre étude actuelle. Il reste cependant de voir de plus près les apports multilingues des deux standards internationaux (récemment imbriquées l'un dans l'autre) désormais considérés comme alternatives incontournables dans l'internationalisation des systèmes d'information et de communication téléinformatique : Unicode et l'ISO/IEC 10646.

3.4.1 L'UNICODE & ISO/IEC 10646

Unicode^[41] est un projet lancé en 1989 pour répondre à la diversité linguistique des systèmes d'information à laquelle l'industrie informatique faisait face. Ce projet était au départ soutenu par les compagnies APPLE et XEROX. Elles ont été rejointes plus tard par Claris, Metaphor, Aldus, Borland, Digital, GO, IBM, Lotus, WordPerfect, Microsoft, NeXT, Novell, Sun, Research Libraries Group avec l'objectif de développer un standard de fait. Ces institutions font actuellement partie du Comité Technique Unicode, le groupe chargé au sein du Consortium Unicode du maintien et de la mise à jour du standard en question. Le terme « Unicode » est utilisé pour la première fois en 1987 par Joe Becker de la compagnie Xerox pour exprimer le concept de « unique, universal and uniform character encoding »^[42]. Ces trois caractéristiques visent l'unicité (aucune duplication des caractères même s'ils apparaissent dans des langues différentes), l'universalité (couverture de toutes les langues écrites modernes) et l'uniformité (tous les caractères ayant le même nombre de bits)^[43].

Pour couvrir les milliers des caractères utilisés dans les textes internationaux, Unicode utilise un jeu de caractères à 16 bits qui lui donne la possibilité d'aller, conformément à la technique de pondération expliquée auparavant, jusqu'à 65.536 combinaisons binaires ($\text{Log}_2 16 = 65.536$).

C'est en 1991 que le groupe de travail informel devient un consortium sous l'acronyme « Unicode, Inc. » qui publia la version imprimée du standard avec ses quelques 27, 000 caractères représentant en priorité la grande majorité des langues de publication du monde moderne.

[40] « Pour limiter le nombre de signes et pour simplifier le traitement des données, l'espace de codage pour les idéogrammes a été codé en utilisant un schéma appelé « unification Han ». Ce travail, une collaboration entre des groupes d'experts des pays directement impliqués et des États-Unis, était soumis aux mêmes règles très strictes que celles utilisées pour les normes japonaises. Cette unification a permis d'éliminer plus de 99 000 signes communs au répertoire des 120 000 signes que compte l'union des jeux de caractères des normes de ces trois langues combinées. De la sorte le nombre de caractères Han a pu être ramené au chiffre indiqué ». Source : Jacques ANDRE et Michel GOOSSENS. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. Op.Cit. p.21.

[41] Quelques parties de la version 2.0 sont disponibles sur l'adresse <http://www.unicode.org/unicode/standard/standard.html>. La différence entre les versions 2.0 et 2.1 sont expliquées sur l'adresse <http://www.unicode.org/unicode/reports/tr8.html>. d'autres rapports techniques sur Unicode peuvent être consultés sur l'adresse <http://www.unicode.org/unicode/reports/techreports.html>.

[42] Jack CAIN. Linguistic diversity, computers and Unicode. <http://www.idrc.ca/library/document/netpac/abs21.html>

[43] Janet C. ERICKSON. Options for Presentation of Multilingual Text : Use of the Unicode Standard. <http://dns.hti.umich.edu/htistaff/pubs/1997/janete.01/>

Jusqu'à 1991, l'Unicode était le seul standard reconnu à échelle internationale pour la codification des caractères sur une matrice de 16 bits. Au même moment, le comité technique ISO (ISO/JTC1/SC2/WG2) développait une première version (Draft) d'une norme ISO structurée sur 32 bits.

Par souci d'unifier cette nouvelle orientation du traitement de l'information multioctets, un accord entre le consortium Unicode et le comité ISO chargé de la norme en question a abouti en 1993 à une fusion des deux projets sous forme d'un jeu de caractères multinational regroupant la majeure partie des caractères utilisés dans le monde : c'est le jeu des caractères universel ISO (Universal Multiple-Octet Coded Character Set) connu sous l'acronyme UCS et normalisé ISO/IEC 10646^[44]. Le standard Unicode devient depuis sa version 1.1 un sous-ensemble 16 bits du jeu des caractères ISO/IEC 10646 à 32 bits dont il occupe désormais le premier plan identifié comme le Plan Multilingue de Base ou *Basic Multilingual Plane* (BMP) (Fig.18 & 19).

Ainsi, les 65,536 points de codes d'Unicode sont devenus les premiers codes de caractères définis par la norme ISO/IEC 10646 auxquels s'ajoutent ceux déjà définis par la norme en question. Les emplacements restants (niveaux supplémentaires) dans les tables à 32 bits de l'ISO/IEC 10646 resteraient vides en prévention d'extensions futures.

Désormais, parler d'Unicode ou de l'ISO/IEC 10646 consisterait en fait à parler du Plan Multilingue de Base (BMP) qui constitue la zone de codes commune entre les deux standards cités. Cette zone est souvent identifiée comme ISO/IEC 10646-2.

3.4.1.1 Aperçu général d'UNICODE

L'un des points clés de la codification sur plusieurs octets est le fait qu'elle attribue des points de codes à des écritures plutôt qu'à des langues. Ce qui veut dire que si plusieurs langues se partagent un ensemble de signes communs, les jeux des caractères propres à chaque langue seraient unifiés dans un seul jeu de caractères communs pour identifier une écriture unique. Cette collection servirait donc à créer un inventaire de symboles utilisés pour écrire plusieurs langues à la fois. C'est le cas des caractères latins (pour les langues latines et germaniques), des caractères arabes (pour les langues arabe, persane, afghane, berbère, ancien turc, swahili ou bantou, malgache, malais, tamoul, ourdou ...)^[45]. Le standard Unicode établit ainsi des codes pour quasi toutes les langues écrites du monde.

Les écritures déjà répertoriées par Unicode dans sa version 2.0 sont par ordre alphabétique : Arabe, Arménien, Bengale, Bopomofo, Cyrillique, Devanagari, Georgien, Grec, Gujarati, Gurmukhi, Han, Hangul, Hebreu, Hiragana, Kannada, Katakana, Latin, Lao, Malayalam, Oriya, Phonetic, Tamil, Telugu, Thai, Tibétain (Voir [Annexe G](#)).

Puisque les caractères Unicode ne spécifient aucunement la langue des textes traités, des protocoles d'un niveau supérieur sont nécessaires pour traiter les exigences linguistiques d'une opération nécessitant la représentation et la désignation de langues. Les opérations de mise en forme des tableaux, des formules mathématiques, de la mise en forme des chaînes de caractères en italique ou en gras, le choix de polices, le balisage de contenu, la structuration des documents, les hyperliens, etc. ressortent du domaine de la représentation enrichie de texte (rich text formats) que des protocoles comme HTML, SGML, Microsoft RTF ... prennent en charge sur la

[44] La référence 10646 est probablement choisie pour commémorer la première appellation utilisée dans le monde des normes internationales pour l'échange des données, en l'occurrence l'ASCII devenu ISO 646.

[45] James G.FEVRIER. Histoire de l'écriture. Paris : Payot, 959.p. 273.

base d'Unicode. Parmi les protocoles de la désignation de langue : les normes ISO/IEC 6429 (Technologies de l'information -- Fonctions de commande pour les jeux des caractères codés), ISO 639 (Codes de représentation des noms des langues), l'avant projet de l'ISO 639-2 à propos de l'extension des codes des langues.

Certains éléments de codes de la BMP ne correspondent pas à des caractères graphiques. Ce sont les codes des caractères de contrôle dont chacun spécifie une fonction particulière sur le texte en phase de traitement (i.e. certains codes de contrôle de la table ASCII : Tabulation, Sonnerie, Stop...).

Parmi les caractères spéciaux de contrôle, la BMP inclut un élément de code qui gère la direction du texte de gauche à droite ou de droite à gauche dans les cas où le système est censé changer de direction d'écriture au niveau d'une seule ligne de texte (§ III-[3.4.2.2](#)).

3.4.1.2 Les techniques de codage sur plusieurs octets et la structure BMP

Le standard Unicode regroupe les caractères en séries d'écritures (scripts)^[46] dans des blocs de codes conformément à leurs caractéristiques d'usage d'origine (i.e. ordre de tri). Les écritures normalisées par les anciennes tables des codes à 7 et 8 bits sont reproduites intégralement dans la structure originale de ces mêmes tables.

La norme ISO/IEC 10646 peut être considérée comme un codage tridimensionnel de caractères, composé de 128 groupes et 256 plans donnant la possibilité de coder 2.147.483.648 caractères différents. Le BMP correspond au plan 0 du groupe 0 (Fig.[20](#) & [21](#)).

Une combinaison binaire ISO/IEC 10646 s'effectue sur quatre octets identifiés dans la nomenclature de la norme, dans un ordre du moins vers le plus significatif, en tant que cellule (octet-C), ligne ou rangée (octet-R), plan (octet-P) et groupe (octet-G) (Voir Fig.[18](#) & [19](#)). La combinaison de ces quatre éléments constitue un caractère unique.

Donc le premier octet d'un code de caractère indique le numéro du groupe; le second octet identifie celui du plan; le troisième et le quatrième spécifient ceux de la rangée et de la cellule respectivement. Les caractères représentés par 2 octets uniquement appartiennent ainsi au plan 0 du groupe 0, en l'occurrence le BMP dans lequel les deux premiers octets d'une représentation sur 4 seraient équivalents à 0.

Unicode est par conséquent une série de tables composées de cellules à valeurs octales et des caractères qu'elles représentent. Le tableau suivant illustre la structure du bloc des caractères Unicode conformes à l'ASCII et à l'ASCII étendu situés entre les valeurs Hexadécimales 0000 et 00FF (Voir la table globale du BMP dans l'[Annexe G](#)).

[46] Une écriture est tout système de représentation graphique par caractères homogènes.

	Contrôle		ASCII							Contrôle		Latin-1					
	000	001	002	003	004	005	006	007	008	009	00A	00B	00C	00D	00E	00F	
0	NUL	DLE	SP	0	@	P	`	p	CTRL	CTRL	NBSP	°	À	Đ	à	đ	
1	SOH	DC1	!	1	A	Q	a	q	CTRL	CTRL	ı	±	Â	Ñ	â	ñ	
2	STX	DC2	"	2	B	R	b	R	CTRL	CTRL	ş	²	Ã	Ō	ã	õ	
3	ETX	DC3	#	3	C	S	c	S	CTRL	CTRL	€	³	Ä	Õ	ä	ó	
4	EOT	DC4	\$	4	D	T	d	T	CTRL	CTRL	¸	´	Å	Ö	å	ö	
5	ENG	NAK	%	5	E	U	e	U	CTRL	CTRL	¥	μ	Ä	Ö	ä	ö	
6	ACK	SYN	&	6	F	V	f	V	CTRL	CTRL		¶	Æ	Ø	æ	ø	
7	BEL	ETB	'	7	G	W	g	W	CTRL	CTRL	§		Ç	×	ç	÷	
8	BS	CAN	(	8	H	X	h	X	CTRL	CTRL	"	.	È	Ø	è	ø	
9	HT	EN	)	9	I	Y	i	Y	CTRL	CTRL	@	±	É	Û	é	ù	
A	LF	SUB	*	:	J	Z	j	Z	CTRL	CTRL	"	°	Ê	Ü	ê	û	
B	VT	ESC	+	;	K	[	k	{	CTRL	CTRL	<	>	Ë	Û	ë	ü	
C	FF	FS	,	<	L	\	l		CTRL	CTRL	~	¼	Ì	Ü	ì	ü	
D	CR	GS	-	0	M	]	m	}	CTRL	CTRL	-	½	Í	Ý	í	ý	
E	SG	RS	.	>	N	^	n	~	CTRL	CTRL	®	¾	Î	Þ	î	þ	
F	SI	US	/	?	O	_	o	DEL	CTRL	CTRL		ı	Ï	ß	ï	ÿ	

Tableau 23 : Le bloc Unicode 0000-00FF

Cette structure est donc bâtie sous forme d'un codage à quatre niveaux hiérarchiques : 128 groupes tridimensionnels (00-7F), chaque groupe se compose de 256 plans bidimensionnels (00-FF), chaque plan dispose de 256 rangées (lignes) unidimensionnelles (00-FF) et chaque ligne est constituée de 256 cellules (00-FF).

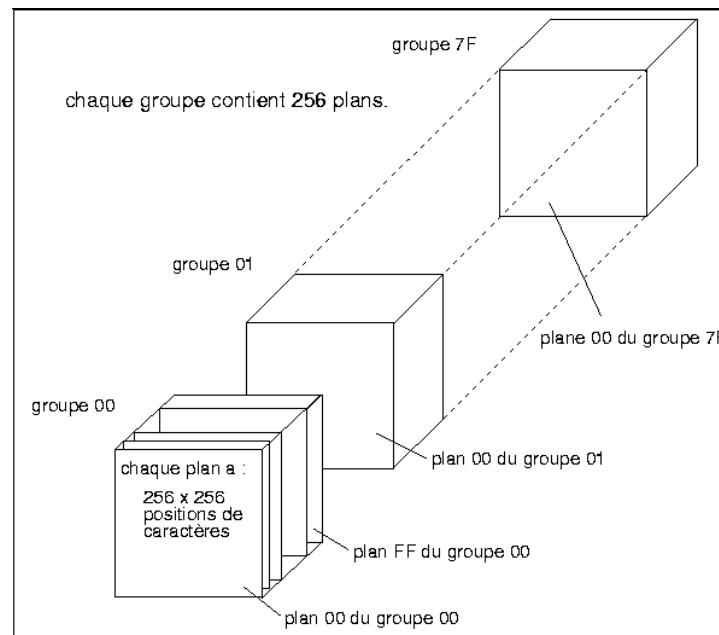


Figure 20 : Structure des espaces de codage ISO/IEC 10646

Source : ANDRE Jacques et GOOSSENS Michel. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. In : Cahiers GUTenberg n°20-mai 1995, pp.1-53

Cette architecture permet de la sorte de coder tous les jeux des caractères du monde entier que la seule BMP (et l'Unicode) ne pourraient couvrir avec les combinaisons binaires à 16 bits.

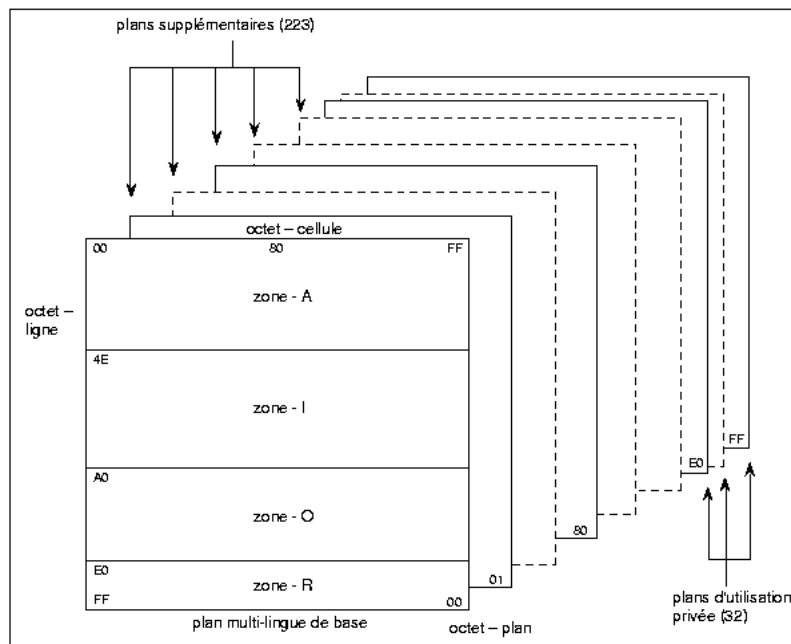


Figure 21 : Structure d'un groupe ISO/IEC 10646

Source : ANDRE Jacques et GOOSSENS Michel. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. In : Cahiers GUTenberg n°20-mai 1995.pp.1-53

Chacune des positions attribuées à un code d'un caractère (aussi appelé point de code) est listée dans les textes de référence en format Hexadécimal précédé du préfixe « U+ ». À chaque caractère, on peut ainsi attribuer 4 octets (G pour groupe, P pour plan, l pour ligne, et C pour cellule). Un octet fait l'objet d'une notation alphanumérique hexadécimale, allant de 00 à FF. Dans ce système, le plan basique multilingue (BMP) est le plan auquel on attribue G = 00 et P = 00. Les caractères y sont donc codés sur 2 octets seulement. Le point de code U+0041 est l'équivalent des valeurs 00 et 41 en hexadécimal (équivalents au nombre décimal 65) qui représentent le caractère « A » dans le standard Unicode (ou BMP).

La répartition des codes au niveau du BMP (qui constitue le seul plan occupé de l'ISO/IEC 10646) est segmentée en quatre zones distinctes réparties entre les rangées 00 et FF ([Fig.20](#)):

- ❑ Une zone A (0000/4DFF) qui réserve 19 903 positions pour les systèmes d'écritures alphabétiques et syllabiques. Les 256 premières positions y sont occupées par les codes ISO 8859-1 (Code ANSI 8 bits) avec l'adjonction d'un octet nul (8 valeurs zéro) pour le convertir en un code 16 bits conformément au codage Unicode. Les positions correspondant aux 32 codes de contrôle ainsi que les codes entre 0x0080 à 0x009F restent vacantes ([Fig.21](#)). Les caractères d'autres écritures occupent les espaces allant entre 0x0000 et 0x2000 avec des interruptions réservées au codage d'autres écritures alphabétiques. Le reste de la zone A est alloué jusqu'au point 0x4000 aux symboles généraux et les caractères phonétiques utilisés en chinois japonais et coréen ;
- ❑ Une zone I (4E00/9EFF) constituée de 20 992 positions réservées à des idéogrammes unifiés CJK utilisés par les langues japonaise, chinoise et coréenne après l'unification Han citée auparavant ;
- ❑ Une zone O (A000/DFFF) qui recense 16 384 positions de codes prévues pour des extensions futures ;
- ❑ Une zone R (E000/FFFD) composée de 8 190 positions dites d'utilisation restreinte pour des usages privés ainsi que pour des caractères spéciaux, des zones de compatibilité et surtout des formes de présentation pour la langue arabe. Elle contient entre autres la zone de codes non définis ailleurs et prévus pour la compatibilité avec d'autres standards.

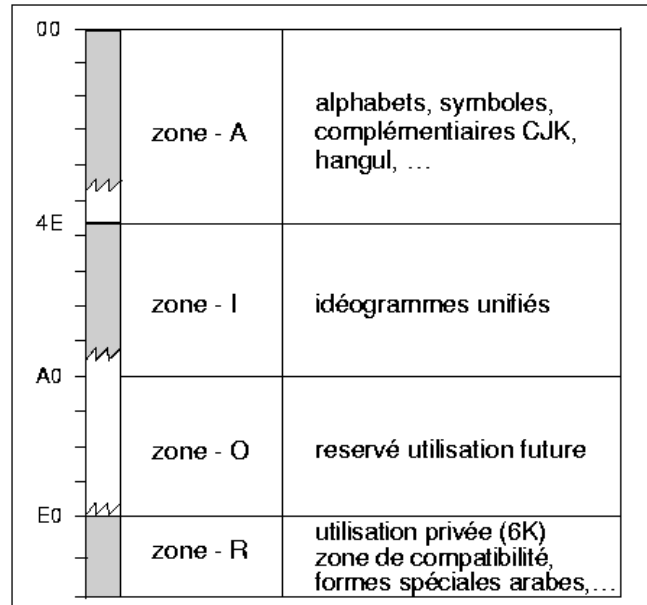


Figure 22 : Les zones du BMP ISO/IEC 10646

Source : ANDRE Jacques et GOOSSENS Michel. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. In : Cahiers GUTenberg n°20-mai 1995, pp.1-53

Dans l'attribution des points de codes au niveau du BMP, trois unités sont à considérer :

- ❑ La cellule ou point de code est équivalente à la valeur du code d'un caractère. ISO/IEC 10646-1 et Unicode 1.1 disposent de 65.536 cellules ;
- ❑ La colonne est composée de 16 cellules (0/9 & A/F en hexadécimal) présentée horizontalement dans la matrice des caractères UCS. ISO/IEC 10646-1 et Unicode 1.1 disposent de 4.096 colonnes ;
- ❑ La rangée, constituée de 256 cellules est visualisée à travers les plans carrés de codes sous forme d'une matrice de 16 cellules par 16 cellules (de 0 à F en horizontal par 0 à F en vertical).

3.4.1.3 Aspects d'usage du codage sur plusieurs octets.

Étant donnée cette complexité dans la structure et dans le codage des caractères UCS, ISO/IEC 10646 a prévu une structuration à trois modes d'usage qui varient en fonction des sous-ensembles des caractères utilisés dans un contexte bien déterminé. Cette structuration est d'ordre à faciliter les opérations d'installation et à éviter la présence de certains sous-ensembles de caractères inutilisables dans des contextes particuliers.

3.4.1.3.1 Les niveaux d'usage du jeu UCS

L'usage des caractères UCS est établi en fonction de trois niveaux dont le BMP constitue la forme la plus complète.

Niveau 1 :

Le premier niveau, et le plus simple, fonctionne à la manière des systèmes traditionnels de codage sur 7 et 8 bits (i.e. ASCII, ISO 8859-1). Chaque caractère graphique occupe une seule

position. Ce modèle est généralement utilisé par les écritures latines, le grec et le cyrillique. À ce niveau, les caractères composés (d'un caractère de base et d'un signe diacritique) sont inclus comme caractères uniques.

Ce niveau inclut également les écritures en arabe et en hébreu qui introduisent une double complexité engendrée d'une part, par la direction de l'écriture de droite à gauche et d'autre part par l'insertion des chaînes de caractères d'une écriture à directionnalité droite-gauche dans une chaîne de caractères d'une écriture à directionnalité opposée (et vice versa). Cette complexité de bidirectionnalité (les incises) est gérée dans ce cas de figure par deux méthodes : à travers la norme ISO/IEC 6429 ou bien par les paramètres Unicode de la gestion de la bidirectionnalité gauche-droite et droite-gauche (U+200E et U+200F respectivement).

Niveau 2 :

Le deuxième niveau d'installation concerne en particulier les écritures de l'Asie du Sud comme le Devanagari. La particularité de ces écritures est qu'elles rendent leur affichage assez complexe puisque dans plusieurs cas, l'apparence et la position des caractères est déterminée par la nature et l'identité des caractères voisins.

Niveau 3 :

Le troisième niveau est le plus complet. Parmi ses caractéristiques :

Une combinaison illimitée de caractères : tous les signes diacritiques peuvent être combinés avec tous les caractères UCS sans aucune limitation. La seule différence avec les autres jeux des caractères codés est que le code du caractère combiné (le diacritique) est sauvegardé après le code du caractère de base et non pas avant. C'est le cas des diacritiques arabes qui, à l'opposé des caractères latins, sont traités après le caractère de base (même dans les normes de codage autres que l'Unicode et ISO/IEC 10646). Un « BA » arabe accentué est une combinaison du caractère « A » suivi du caractère non espacé « Fatha » alors que le caractère « ê » (e avec un circonflexe dessus) est le signe non espacé « ^ » suivi du caractère de base « e ».

3.4.1.3.2 La compatibilité inter-normes

La compatibilité avec les formes de codage en vigueur était l'un des plus importants défis au nouveau standard. Le passage des applications en cours d'un système de codage à un autre aurait évoqué une grande polémique d'acceptation ou de rejet si un degré de transparence dans la migration vers le nouveau code n'avait pas été garanti. Déjà l'histoire des techniques du codage des caractères ne cesse de connaître des divergences entre l'ASCII et le codage étendu du groupe ISO/IEC 8859. La polémique se situe en réalité au niveau du mécanisme de codage et d'interprétation des codes utilisés au moment de la rédaction ou de la réception d'un corpus textuel. Le rôle d'Unicode est défini, comme l'a été l'ASCII et les tables des pages de codes, au niveau de cette jonction entre le code du caractère et son rendu visuel.

Face à un utilisateur qui saisit du texte à partir d'un clavier, le système en cours d'utilisation identifie la touche sur laquelle cet utilisateur a appuyé et procède en conséquence à déchiffrer la combinaison binaire du caractère concerné. Une lettre « T » codée en Unicode U+0054 (hexadécimal) sera mémorisée en mémoire pour être ensuite traitée par le programme qui gère l'affichage. Ce dernier, qui pourrait être à titre d'exemple un gestionnaire Windows ou un module particulier de l'application en cours, se chargera d'établir le rapport entre les codes des caractères indexés et l'image des caractères qui leur correspondent pour les afficher sur écran.

C'est là un des principes les plus importants d'Unicode : n'adresser que le codage et l'interprétation du texte exclusivement. Dans le cas où un utilisateur emploierait un programme de traitement de texte, la vérification de l'orthographe ou la gestion des retours chariot resteraient en dehors des compétences techniques d'Unicode qui se contenterait d'accomplir la phase du codage et de l'interprétation des codes. Dans la majorité des cas et en fonction des méthodes qu'Unicode utilise pour coder les caractères (il y en a plusieurs présentées ci-après), un fichier texte ASCII converti en un fichier texte Unicode (Unicode Plain Text) aurait une taille double partant du principe du codage sur 2 octets de chaque point de code. Par contre le même fichier ASCII mis en forme par un programme de traitement de texte (ou toute autre application de Bureautique ou de traitement spécial), donc constitué des codes des caractères en plus des codes de formatage, aurait une taille augmentée de 20 à 25 pour cent car seules les données textuelles seraient doublées par une conversion Unicode.

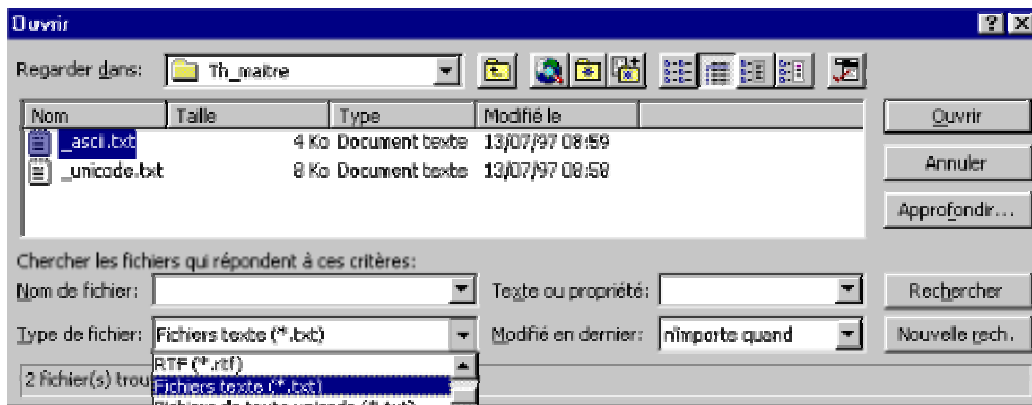


Figure 23 : Dédoublage de taille pour un même fichier du format ASCII au format Unicode

Pour la majeure partie des applications qui se basaient sur un codage ASCII et Latin-1, aucun problème majeur ne se posait. Les combinaisons binaires de ces deux normes (à 7 ou à 8 bits) figurent en tête de file dans les premiers blocs de codes de la BMP. La majorité des autres normes en vigueur peut aussi être traduite (mapped) puisque Unicode inclut la totalité des caractères définis par ces normes en question.

En effet, la définition d'Unicode a respecté depuis sa création en 1991 un degré de compatibilité avec les normes précédentes de sorte qu'un texte conçu sous les anciennes normes peut facilement être converti en Unicode. Les normes compatibles avec Unicode sont illustrées dans le tableau suivant.

ANSI X3.4	GB 2312	ISO 5428	ISO 9036
ANSI X3.32	ISCII	ISO 6438	ISO/IEC DIS 10367
ANSI Y10.20	ISO 646	ISO DIS 6861.2	ISO/IEC DIS 10646
ANSI Z39.47	ISO 2047	ISO DIS 6862.2	JIS X 0208
ANSI Z39.64	ISO 2022	ISO 6937	JIS X 0212
ASMO 449	ISO 2033	ISO 8859	KS C 5601
CCCI	ISO 5426	ISO 8879	TIS 620-2529
CNS 11643	ISO 5427	ISO DP 8957	

Tableau 24 : Normes et standards compatibles avec Unicode

Le tableau suivant démontre également le degré de compatibilité entre plusieurs techniques de codage à un ou plusieurs octets. Il illustre parfaitement le maillage entre ASCII, ISO Latin-1,

Unicode et ISO/IEC 10646 par l'adjonction de valeurs « 0 » initiales selon la structure binaire du code concerné.

Bits	Standard	Valeur binaire	Hex.	Dec.	Car.
7	ASCII	1000001	41	65	A
8	ISO 8859-1 (Latin-1)	01000001	41	65	A
16	Unicode (10646 UCS-2) (10646 BMP)	00000000 01000001	41	65	A
32	10646 UCS-4	00000000 00000000 00000000 01000001	41	65	A

Tableau 25 : La valeur binaire du caractère A selon plusieurs standards

Avec l'unification d'Unicode et de l'ISO/IEC 10646, la question était de savoir laquelle des normes faudrait il utiliser. Transmettre des données sur 4 octets n'est point pratique car beaucoup d'applications en usage ne parviennent même pas à utiliser les 16 bits (2 octets) d'Unicode. Plusieurs protocoles continuent à traiter les octets entre les valeurs hexadécimales 00 et 7F (128 combinaisons). Il est d'ailleurs très courant de voir des applications de messagerie électronique, à l'instar du protocole SMTP, interdire les caractères au-delà des 128 positions supérieures de l'ISO 8859-1.

Pour remédier à cette diversité dans les modes de codifications des caractères, plusieurs méthodes de transformation algorithmiques des données en UCS ont déjà été définies pour les opérations de transmission et d'échange : UTF-7, UTF-8 et UTF-16 qui accompagnent les versions canoniques UCS-2 (Unicode) et UCS-4 (10646). Le terme UTF est l'acronyme du concept « UCS Transformation Format » suivi du nombre de bits admis par le canal de transmission. La possibilité de traduction inter-normes a ainsi élargi le champ d'action du codage à plusieurs octets et par conséquent renforcé son adaptation par un plus grand nombre de fabricants de matériels et de concepteurs de logiciels pour assurer une transition graduelle et sans rupture vers le tout Unicode.

- **Format UTF-1**

La méthode UTF-1 (UCS Transformation Format No.1) définie dans un annexe de l'ISO/IEC 10646, n'est plus valide, donc obsolète et retirée du contexte d'usage de la dite norme.

- **Format de transformation 7-Bits : UTF-7**

L'avantage de UTF-7 est sa capacité de coder tout le répertoire Unicode en utilisant les octets à 7 bits de l'US-ASCII. Cette transformation à 7-bits est prévue pour une utilisation avec la norme MIME et d'autres supports de transmission à sept bits (Messagerie électronique, News) utilisant le protocole SMTP. UTF-7 transforme un texte codé en UCS-2 en une séquence d'octets ne dépassant pas la position 7F (128 points de codes). Il devrait être utilisé uniquement dans des contextes de transfert de données codées à 7 bits. Les caractères au-delà des 128 premiers codes y sont accessibles à travers les séquences d'échappement. Dans d'autres contextes, UCS-2 ou UTF-8 sont plus recommandés.

- **En format de transformation 8-Bits : UTF-8**

Le format UTF-8 pour le codage des données Unicode (et ISO/IEC 10646) était initialement proposé en septembre 1992 par Ken Thompson de la compagnie AT&T Bell Laboratories en collaboration avec XOJIG (*X/Open Joint Internationalization Group*). Il était connu sous le nom

de FSS-UTF « *File System Safe Universal Character Set Transformation Format* » avant de devenir UTF-8^[47].

UTF-8 transforme un flux de données codées en Unicode (16 bits) ou ISO/IEC10646 (32 bits) en une combinaison à 8 bits conforme à la première moitié de la première rangée du BMP. Cette méthode a l'avantage de préserver les mêmes codes des caractères ASCII dans la table Unicode entre les rangées 00 et 7F. Chacun de ces codes est représenté par un octet qui sauvegarde l'unicité de la correspondance entre la valeur de l'octet en question et le caractère US-ASCII correspondant.

Les autres codes UCS sont automatiquement transformés entre deux et six octets dans la rangée 80-FF. Ainsi UTF-8 peut être facilement traduit en UCS-4 sans l'obligation de subir une conversion en UTF-16. Il peut également être utilisé par un grand nombre d'applications sans nécessité de modifications profondes dans la réécriture de leurs codes sources.

UTF-8 codifie les caractères UCS-2 ou UCS-4 dans un nombre variable d'octets qui dépend de la valeur de chaque caractère qui lui est attribuée par la norme ISO/IEC 10646. Les caractéristiques de ce format de transformation sont précisées par François Yergeau comme suit^[48] :

- ❑ Les valeurs des caractères entre 0000 0000 et 0000 007F (répertoire US-ASCII) correspondent aux octets entre 00 et 7F (les valeurs 7 bits de l'US-ASCII) ;
- ❑ Ces valeurs de l'US-ASCII ne sont attribuées à aucun autre caractère dans le filtre de transformation UTF-8. Ceci est d'ordre à garantir une compatibilité avec les systèmes de gestion de fichiers ou tout autre système qui fonctionne sur la base de l'ASCII tout en restant transparent pour les autres valeurs (Gestionnaire de fichiers, la fonction « printf() » dans les bibliothèques C) ;
- ❑ Une conversion à double sens est possible entre UTF-8 et UCS-4 et UCS-2 ;
- ❑ Le premier octet d'une séquence multioctets indique le nombre d'octets dans la séquence (Tab.26) ;
- ❑ Les limites des caractères sont facilement repérables n'importe où dans la séquence d'octets ;
- ❑ L'ordre de tri alphabétique des séquences codées en UCS-4 est préservé bien que l'élément de tri soit d'une importance seconde puisqu'il n'a aucune valeur culturelle ;
- ❑ Les valeurs des octets FE et FF ne sont pas impliquées.

Nous incluons ci-après un exemple du mécanisme de transformation et de codage entre UCS4 et UTF-8 emprunté à François Yergeau^[49].

En UTF-8, les caractères sont codés en utilisant des séquences de 1 à 6 octets. Une séquence d'un seul octet aurait la valeur du bit le plus important de l'octet en question égale à « 0 ». Le reste des 7 bits de l'octet sont utilisés pour indiquer la valeur du caractère.

Dans une séquence de « n » octets où « n > 1 », l'octet initial (le premier à gauche) aurait les « n » bits les plus importants (à gauche) égales à « 1 » suivi par un bit équivalent à « 0 ». Le reste des bits de cet octet initial indiquera les valeurs binaires du caractère à coder. Tous les octets suivants de la séquence auront les deux bits les plus importants mis aux valeurs « 1 » et « 0 »

[47] UTF-8, a Transformation Format of Unicode and ISO 10646. <http://www.imc.org/rfc2279>

[48] F. YERGEAU. UTF-8, a transformation format of Unicode and ISO 10646. RFC 2044. Network Working Group. October 1996 (<http://sunsite.auc.dk/RFC/rfc2044.html>)

[49] F. YERGEAU. UTF-8, a transformation format of Unicode and ISO 10646. Op.Cit.

respectivement. Les six (6) bits qui restent dans chacun des octets indiquent les valeurs des caractères à coder.

Le tableau suivant résume le format de ces différents types d'octets dans une transformation UCS-4 / UTF-8. La lettre « X » indique les bits qui correspondent à la valeur des caractères en UCS-4.

UCS-4 (en hexadécimal)	Séquence d'octets en UTF-8 (en binaire)
0000 0000-0000 007F	0xxxxxxx
0000 0080-0000 07FF	110xxxxx 10xxxxxx
0000 0800-0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000-001F FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
0020 0000-03FF FFFF	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
0400 0000-7FFF FFFF	1111110x 10xxxxxx ... 10xxxxxx

Tableau 26 : Transformation UCS-4 / UTF-8

Source : F. Yergeau. UTF-8, a transformation format of Unicode and ISO 10646. RFC 2044. Network Working Group. Octobre 1996 (<http://sunsite.auc.dk/RFC/rfc/rfc2044.html>)

Le consortium Unicode encourage l'usage de UTF-8 comme première étape de l'introduction de sa norme de codage à 16 bits. D'ailleurs la conversion UTF-8/UTF-16 s'effectue actuellement sans perte de données (voir le RFC^[50] 2044 du Network Working Group par F.Yergeau à l'URL <http://sunsite.auc.dk/RFC/rfc/rfc2044.html> pour plus de détail).

• Format canonique étendu : UTF-16

Une transformation pour obtenir UCS-4 à l'aide de codes UCS-2. Cette extension utilise deux fois 1024 codes UCS-2 pour représenter un million de caractères additionnels. Ce format est prévu pour coder tous les caractères ISO/IEC 10646 qui n'avaient pas de place dans le BMP (Hiéroglyphes, Chinois rare ...).

UTF-16 réserve 1024 codes dans le BMP pour la moitié supérieure des codes et 1024 autres pour la moitié inférieure du BMP. Quand ces codes sont combinés, ils permettent de rajouter 917,504 positions de codes pour des caractères supplémentaires et 131,072 espaces supplémentaires pour un usage privé.

La motivation derrière le développement de ce format est de garantir aux développeurs d'applications en Unicode une compatibilité avec les extensions futures des caractères UCS en dehors des limites du BMP.

Ces formats de transformation sont très importants pour résoudre des problèmes de compatibilité entre les formats Unicode ou ISO/IEC 10646 et leurs prédécesseurs à 7 et 8 bits. UTF-7 et plus particulièrement UTF-8, très utilisés dans la communauté informatique, peuvent servir pour la conversion à partir de et vers Unicode et ISO/IEC 10646. La compatibilité entre UTF-8, utilisé dans Microsoft Office 97 et UTF-16 utilisé par Unicode dans Windows NT crée un point d'accès entre les deux systèmes d'exploitation.

[50] Request For Comment : Etape préliminaire du passage d'une proposition de norme à une norme définitive.

3.4.1.3.3 Des handicaps d'usage

Ainsi défini, Unicode reste pourtant limité au codage des caractères des langues du monde. Il ne pourra en aucun cas être considéré comme :

- ❑ Un système de tri. Le tri est traité au niveau applicatif ;
- ❑ Une définition de glyphes ;
- ❑ Un système de comparaison de chaînes de caractères ;
- ❑ Un système de formatage de texte ;
- ❑ Un identifiant de langue ;
- ❑ Un mécanisme de conversion de caractères.

Puisque toutes ces fonctions demeurent toujours indispensables dans le traitement des documents textuels, chacune dispose d'une norme correspondante et même de plusieurs à la fois. Ceci représente en perspective un travail de conciliation considérable pour permettre finalement à un grand nombre de règles et de standards de définir de manière adéquate le contexte de travail de l'utilisateur.

Le passage progressif des applications du domaine public et commercial à l'univers Unicode (i.e. Unix et Windows NT surtout) va instaurer une nouvelle routine d'identification des langues dans le développement des applications futures. Ceci n'empêcherait pas que l'adaptabilité d'Unicode au nouveau contexte informationnel nécessiterait d'une part la transition massive des concepteurs informatiques (logiciel/matériel) vers le multioctets de base, et d'autre part la réhabilitation des applications précédentes au nouveau standard en vogue.

Déjà, des problèmes majeurs dans les attributions même d'Unicode restent encore en quête de solutions :

- ❑ **Les écritures locales** : Les caractères des langues locales comme le chinois, le coréen permettent l'invention de nouveaux caractères pour des usages locaux. Ceci est d'ordre à accentuer l'ambiguïté de la représentation des caractères et de fausser la caractéristique fondamentale de l'unicité de codage adoptée par Unicode.
- ❑ **Les langues non couvertes** : dans le cas où une langue n'est pas incluse dans la zone des codes, Unicode ne serait d'aucun intérêt à moins d'utiliser les zones privées pour y inclure un codage particulier.
- ❑ **La conversion des données** : La concordance entre Unicode et les systèmes locaux passe par des procédures assez longues pour la création des tables de correspondances et la conversion des données vers et à partir d'Unicode.
- ❑ **Affichage et impression des données** : Certains systèmes d'une renommée et d'une envergure internationale, comme MS-DOS et Windows 3.1, ne reconnaissent pas le codage Unicode. Des procédures de conversion spéciales sont encore à définir pour permettre la double transparence entre ces systèmes.
- ❑ **La saisie des données** : Unicode n'est pas très bien adapté à la saisie des données. Au contraire, il compliquerait davantage la situation. Un utilisateur travaillant sur un système Unicode est supposé effectuer sans effort, la saisie de tous les caractères disponibles dans le dit système. Or ceci n'est pas facile sans entraînement préalable de longue haleine.

- ❑ **Le tri** : Unicode ne résout pas le problème du tri qui devrait à notre sens rester l'une des attributions de la localisation des systèmes concernés. D'ailleurs, le tri minimum accompli à travers les tables des codes à 7 et à 8 bits précédentes par l'ordre triée et le positionnement contiguë des caractères codés n'a plus de justification dans une table de codes unifiés qui n'applique pas les techniques d'extension de codes à la manière d'ISO/IEC 2022. Les différences linguistiques, sémantiques et culturelles entre les langues d'une même écriture obligeraient le recours à la mise en place de tables de tri locales. Une même langue pourrait même faire objet d'une divergence de méthode de tri. Nous traiteront le cas du tri de la langue arabe dans le chapitre suivant.
- ❑ **La localisation** : Unicode ne traite que les codes des chaînes des caractères dans un format texte. Or, un document structuré et destiné à la transmission, a besoin d'un plus d'information de formatage (i.e. SGML^[51], RTF^[52]...) et de désignation de contenu (i.e. MARC^[53], CCF^[54]...) pour que le système récepteur puisse le déchiffrer convenablement.

Unicode constituerait, selon Peter L. Noerr^[55], un pas considérable vers l'internationalisation des systèmes d'échange de l'information bien qu'il ne constitue qu'une partie de la solution définitive. Le transcodage unifié du corpus textuel échangé représenterait selon ses estimations autour de 20% du problème global du handicap linguistique imposé aux systèmes d'échange internationaux. Un facteur aussi déterminant dans la généralisation de l'usage Unicode est sans doute le poids que représente le contexte américain. Malgré les avantages évidents qu'une plateforme Unicode représente, Janet Erickson^[56] estime que son introduction aux États Unis d'Amérique reste assez lente. La majorité des utilisateurs américains n'exprime pas un grand besoin d'aller au delà des 256 points de codes de Windows qui répondent largement à leurs propres besoins linguistiques.

3.4.1.3.4 Unicode et le traitement des données textuelles

Le bilan de l'internationalisation des systèmes de codage de l'information resterait à notre sens très positif. La codification sur plusieurs octets est devenue un choix incontournable pour les développeurs d'applications et les constructeurs de matériels informatiques. Déjà le monde de l'industrie des nouvelles technologies de l'information a suivi ce courant de multilinguisme dans le traitement des données. Le consortium Unicode est constitué dans sa majorité d'industriels de grande réputation internationale. L'adoption du nouveau standard est aussi entreprise par les grands gourous du monde du génie logiciel.

En effet, Unicode/10646 et les formats UTF ont été bien accueillis dans les univers informatiques et informationnels comme solution aux problèmes de l'internationalisation des plates-formes de travail. Unicode est progressivement appréhendé comme le jeu de caractères universels destiné à remplacer l'ASCII ainsi que toutes autres formes de codage mono ou multi-octets en usage. Déjà les grands producteurs de logiciels comme Microsoft, IBM et Apple ont commencé depuis 1991 à intégrer la solution Unicode dans les dernières versions de leurs produits. Microsoft, étant donnée l'universalité de ses produits, est un exemple très prometteur. Windows NT 4.0 est entièrement basé sur un codage Unicode à 16 bits. Windows 95 est moins

[51] Standardized General Markup Language

[52] Rich Text Format

[53] MACHine Readable Cataloguing

[54] Common Communication Format

[55] Peter L. NOERR. Information Management and Engineering Ltd. London. Voir à l'adresse [http :
http://www.ua.ac.be/KB/pn/pnoerr6.html](http://www.ua.ac.be/KB/pn/pnoerr6.html)

[56] Janet C. Erickson. Options for Presentation of Multilingual Text : Use of the Unicode Standard. <http://dns.hti.umich.edu/htistaff/pubs/1997/janete.01/>

performant dans ce sens car il part essentiellement d'une conformité avec les codes Windows 3.x et les jeux des caractères locaux pour le traitement interne des données.

Les applications qui tournent sous Windows 95 peuvent pourtant utiliser Unicode si les chaînes de caractères reçues sont converties via le jeu local de caractères avant d'être traitées en interne. Sous Windows NT, les programmes qui utilisent les jeux des caractères locaux au lieu des jeux Unicode nécessitent aussi une étape de prétraitement par conversion entre jeux de caractères. MS Word 97 permet déjà de sauvegarder des fichiers en format UTF-8 et en format HTML.

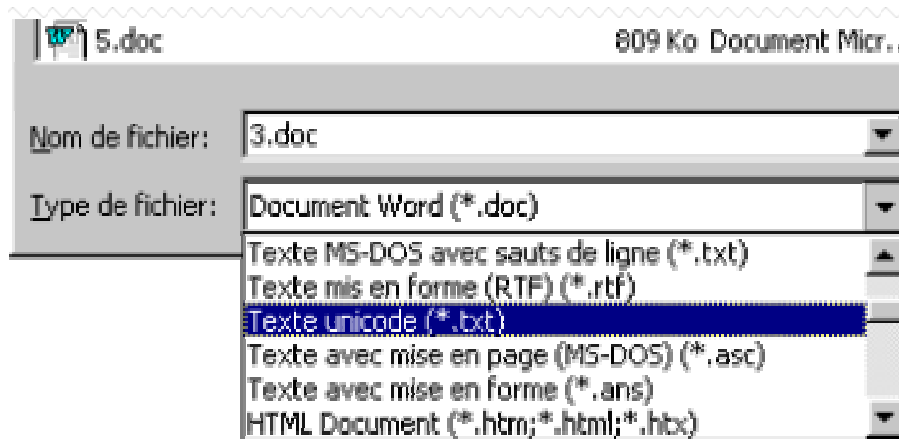


Figure 24: Sauvegarde en format Unicode sous MS Word 97

C'est aussi le cas de l'éditeur de page Web Microsoft FrontPage et de l'utilitaire Notepad qui permettent une sauvegarde en format UTF-8. UTF-8 est aussi disponible sous Netscape Navigator 3.0 sous plates-formes Windows 95 et NT 4.0. Le navigateur Spyglass Mosaic supporte également UTF-8 sous Windows 95 à l'instar des navigateurs Tango d'Alis et du navigateur « Navigate With Accent » d'AccentSoft conçus tous les deux selon la technologie Spyglass. Nous reprendrons ces détails dans le point 4 du présent chapitre pour couvrir l'aspect multilingue des solutions Internet (navigateurs multilingues, éditeurs des pages Web, gestionnaires de messagerie etc.) dans un environnement partagé sous un système ouvert.

L'univers Macintosh n'applique pas encore le standard Unicode dans les opérations élémentaires du système 7. Son implication est prévue avec la version Système 8.

En plus des systèmes d'exploitation, la consolidation de la conversion Unicode passe aussi par la préparation des fontes et des systèmes de fontes adéquates avec la norme en question.

La différence entre l'identification des valeurs des codes et les formes de rendu visuel ou imprimé (glyphes) est cruciale pour comprendre le rôle d'Unicode dans le traitement des séquences textuelles. Le caractère comme identifié par Unicode est en réalité une entité abstraite reconnue par son nom de code i.e. « LATIN CHARACTER CAPITAL A » ou « ARABIC DIGIT 5 ». La forme du rendu est la représentation visuelle du caractère.

Il est peut être important de rappeler à ce stade qu'une séquence d'octets pourrait être interprétée par une multitude de façons quand elle est convertie en caractères (ou autre chose !). Il est en effet difficile de distinguer dans une séquence d'octets l'équivalent et les limites réelles de chaque octet. Il est même difficile de savoir si l'octet représente un seul caractère ou s'il fait partie d'une présentation d'un caractère à deux octets. Il faudrait également prévoir qu'une séquence d'octets peut être l'équivalent codé d'une image en format « Bitmap » ou d'un fichier

programme en format « Binaire » ou tout simplement une donnée numérique traitée par l'ordinateur dans un format interne.

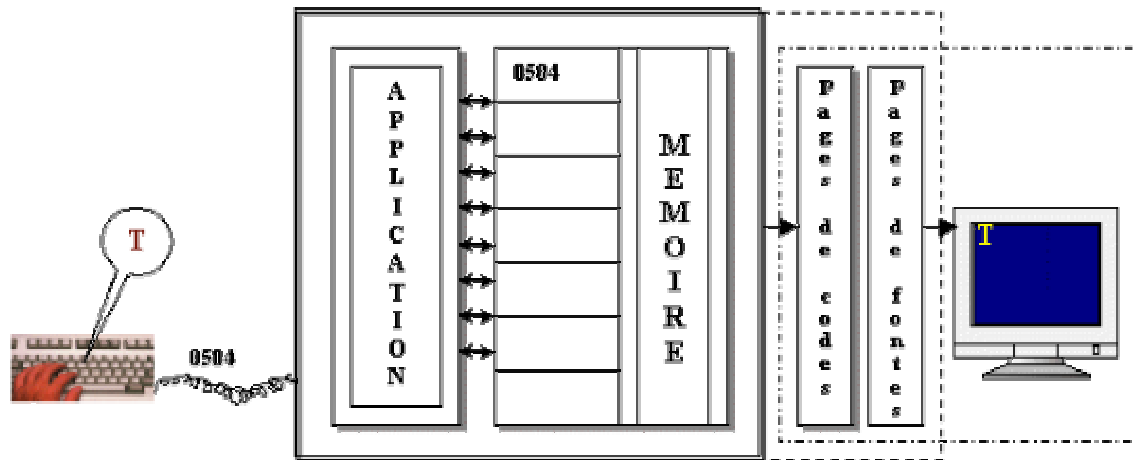


Figure 25 : Mode de conversion code/glyphe par pages de codes et de fonte

Par conséquent, le standard Unicode définit uniquement les codes de caractères. Il ne résout pas les problèmes liés au tri, à la sélection des glyphes pour les rendus visuel et imprimé ainsi qu'à d'autres manipulations textuelles de haut niveau. Ces opérations restent tributaires des applications et des couches systèmes qui lui sont superposés. Pour toutes ces applications multi-octets, des fontes compatibles avec Unicode sont nécessaires.

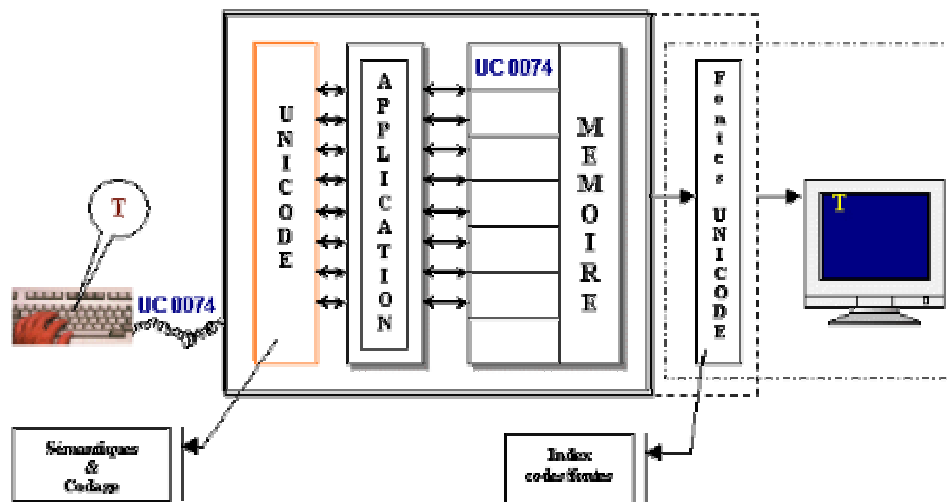


Figure 26 : Mode de conversion code/glyphe par fontes Unicode

Puisque le codage sous Unicode se base essentiellement sur les valeurs des codes des anciennes normes, un moyen de conformité de conversion vers Unicode est assuré. Tous les caractères identiques en forme appartenant à des normes différentes, sont identifiés et traduits en un seul caractère Unicode.

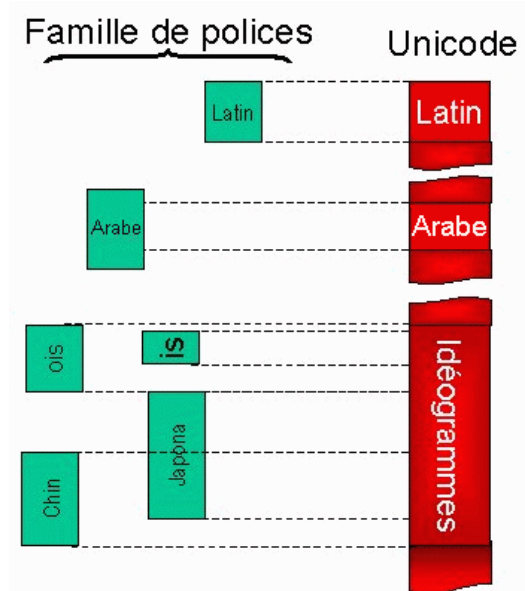


Figure 27 : Unification des fontes Unicode

Des solutions d'installation et d'usage d'Unicode commencent à proliférer. Certaines sont partielles et procèdent par tables de traductions entre modules hétérogènes. D'autres sont intégralement basées sur une gestion des caractères multi-octets.

Parmi la longue liste évolutive d'applications et de matériels conçus à base d'Unicode, se trouvent en tête de file^[57] :

- ❑ SP: c'est un compilateur SGML de la deuxième génération qui fonctionne avec des jeux de caractères codés à seize bits. Son gestionnaire des entités accepte les codages UTF-8 et UCS-2 (Unicode) ;
- ❑ La société *Progress First Software* vend des polices de caractères PostScript Typographic International Series compatibles avec Unicode ;
- ❑ *Windows NT* de Microsoft : Si *Windows 95* au niveau de son système d'aide fonctionne avec Unicode, la version NT fonctionne directement en 16 bits et dispose d'une police « Lucida Sans Unicode » qui contient quelque 1752 symboles ;
- ❑ Y & Y Inc. : Commercialisent des polices LATIN-2, Cyrillique et Unicode^[58] ;
- ❑ *X/Open* s'est orienté vers l'internationalisation des logiciels par l'intégration d'un support des différentes langues ;
- ❑ Le système plan 9 d'AT&T offre une plate-forme d'entrée pour Unicode ;
- ❑ *Gamma Production* commercialise plusieurs logiciels capables d'intégrer le traitement multilingue (toutes les langues Unicode) à l'intérieur même de *Windows 3.1* ou NT (*Gamma Server for Unicode*). *Gamma Unitype* est un utilitaire supplémentaire capable de rendre possible la saisie directe de plus de 175 langues différentes sous *Windows 3.1* ou NT. *Gamma Universe Windows* est aussi un système multilingue Unicode de traitement de texte sous *Windows* avec des polices pour traiter la plupart des langues du monde. Toutes ces applications sont munies d'outils linguistiques pour la vérification d'orthographe, et le traitement complexes des ligatures ;
- ❑ Le Système C3 développé par l'Initiative Trans-Européenne (TERENA) permet la traduction (mapping) entre les différents jeux de caractères^[59] ;

[57] Source : Jacques ANDRE et Michel GOOSSENS. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. Op.Cit. p.25.

[58] Voir leur page d'accueil WWW à l'URL <http://www.yandy.com/>

- ❑ MASS (Multilingual Application Support Service) est un logiciel de développement d'applications multilingues mis en place par « Institute of Systems Science » à Singapour à bases de plusieurs variantes linguistiques pour le codage et la saisie des données^[60]. C'est un jeu d'outils logiciels qui permet de gérer un environnement multilingue pour les utilisateurs finals ainsi que pour les développeurs de programmes. Il dispose d'un éditeur de texte basé sur Unicode, d'un émulateur pour X-terminal universel et d'une série de bibliothèques du langage C qui contiennent des Interfaces de Programmation d'Applications (Application Programming Interface : API)^[61] nécessaires pour le développement d'applications multilingues.
- ❑ π , une extension 16-bit de *T E X* utilise Unicode comme codage interne ;
- ❑ L'une des versions récentes de *Mule*, l'extension 16-bits multilingue de *Gnu emacs* développée au Japon, ajoutera Unicode à sa liste de codages possibles pour éditer des textes ;
- ❑ SGML (Standardized General Markup Language) prévoit également le développement d'une syntaxe concrète de référence étendue. Cette proposition ERCS (Extended Reference Concrete Syntax) prévoit l'usage des balises codées en ISO/IEC10646 en plus de l'ASCII courant. De la sorte, « les utilisateurs des pays de l'Asie du sud-est devraient être capables de saisir, éditer, baliser, sauvegarder et envoyer leurs documents SGML dans leur propre langue en utilisant un système de codage de leur choix »^[62] ;
- ❑ L'univers du World Wide Web connaît aussi un grand nombre d'initiatives avec des solutions qui visent un rendu de textes codés en plusieurs normes. « Récemment un collaborateur de la Commission des Communautés européennes à Bruxelles a proposé un certain nombre d'extension multilingues aux clients WWW : Unicode deviendrait l'alphabet de base du Web et pour « savoir » quelle est la langue dans laquelle un document donné est rédigé, il introduit la notion d'hypertexte multilingue aligné (MAH, Multilingual Aligned Hypertext) qui permet un accès transparent à la version (langue) désirée en y associant une « étiquette » codée d'après la norme ISO 639... »^[63].

Puisque Unicode est la réplique intégrale de l'ISO/IEC 10646 (au niveau de la BMP), il reprend chaque caractère ISO/IE 10646 selon son code et son adresse d'origine. Cependant il rajoute à cette exactitude d'adresse une série de contraintes qu'il impose sur la forme du codage adopté, les niveaux de leur usage ainsi que leur façon d'agir dans des contextes particuliers de traitement textuel. Ainsi l'adoption unanime du codage sur plusieurs octets n'a pas dissipé définitivement certaines différences fondamentales qui subsistent encore entre Unicode et ISO/IEC 10646. Les particularités d'usage d'Unicode peuvent être résumées en ce qui suit :

- ❑ Unicode est d'abord beaucoup plus simple que ISO 10646 ;
- ❑ Cependant, il n'offre pas les mêmes avantages et possibilités de présentation ;
- ❑ De plus, Unicode est prévu pour une utilisation à court terme, contrairement à la norme ISO 10646 qui a des objectifs plus lointains ;
- ❑ Unicode est un code à 16 bits uniquement alors qu'ISO/IEC 10646 définit des points de codes à 2 octets (UCS-2) et à 4 octets (UCS-4) à la fois ;
- ❑ Par la nature de son envergure, ISO/IEC 10646 est limité aux simples aspects de codage dans les différents standards. Unicode inclut par contre des spécificités additionnelles pour le processus de son installation ;

[59] Plus d'information est à l'URL <http://www.nada.kth.se/i18n/c3/>

[60] Voir l'URL <http://www.iss.nus.sg/RND/MLP/Projects/MASS/MASS.html>.

[61] L'API est une interface normalisée par l'intermédiaire de laquelle le programmeur a directement accès aux fonctions d'un système d'exploitation ou d'une interface utilisateur (par exemple Microsoft Windows).

[62] Voir ERCS à l'URL : <http://www.sgmlopen.org/sgml/docs/ercs/ercs-home.html>

[63] Voir l'URL <http://www.echo.lu/other/norm>

- ❑ Unicode définit les valeurs sémantiques des caractères de manière plus explicite qu'ISO/IEC 10646. Il précise, à titre d'exemple, l'ordre d'affichage d'un document multilingue et bidirectionnel (du texte arabe avec des chiffres ou une incise arabe dans une suite textuelle latine). Nous aborderons une partie de cet aspect de bidirectionnalité dans les chapitres suivants ;
- ❑ Unicode prévoit une série de tables d'attributs de caractères et de conversion vers d'autres jeux de caractères ;
- ❑ Unicode fait appel à un algorithme complet de bidirectionnalité pour le traitement de textes multilingues comme l'arabe ou l'hébreu avec indication des caractères de formatage spécifiques pour ce genre de traitement. ISO/IEC 10646, bien qu'elle couvre les caractères de formatage utilisés par Unicode, elle ne spécifie pas le type d'algorithme qui les appliquera pour la présentation adéquate de la bidirectionnalité ;
- ❑ Unicode spécifie la forme des opérations de tri et fait usage des double diacritiques non espacés pour la formation des caractères composés (Overstiking);
- ❑ Unicode effectue une traduction (conversion) des tables de caractères compatibles (nationales et internationales) vers des tables normalisées Unicode ;
- ❑ Unicode définit le comportement graphique de certaines écritures cursives comme l'arabe ou le Devanagari au niveau des ligatures, superimpressions, analyse contextuelle...
- ❑ Unicode dispose d'un algorithme pour déterminer l'équivalence de deux chaînes de caractères contenant des diacritiques non espacés. Il permet aussi de décomposer de façon univoque un caractère précomposé. Cet algorithme d'analyse de glyphe lui permet de distinguer, dans une combinaison précomposée, entre les caractères espacés et non espacés ;

Unicode a été conçu essentiellement pour répondre aux besoins d'applications multilingues mixant des langues latines et des langues extrêmes orientales (chinois, japonais, coréen, etc.) et ne développe pas beaucoup la partie relative à la langue arabe. Par contre ISO 60464 a une portée plus large et doit donc contenir, à terme, les codes des caractères de tous les alphabets existants et pour lesquels une normalisation est envisagée, ainsi que l'ensemble des variantes et spécificités de toutes les langues issues de ces alphabets.

3.4.2 La langue arabe et le BMP :

Depuis la première publication de la norme ISO/IEC 10646 en 1989, le groupe de travail arabe (ATG) auprès de l'ECMA^[64] a présenté un premier projet d'une codification du caractère arabe qui s'adapterait vaguement aux besoins de codification sur plusieurs octets. La dissolution de l'ASMO en 1990 et par conséquent, la suppression des activités de la commission du caractère arabe, aucune contribution ultérieure n'a été fournie pour promouvoir la présence du caractère arabe dans ce projet universel. L'état actuel de la présence du caractère arabe dans le BMP est issu d'une contribution d'une commission égyptienne qui a pu introduire les éléments suivants dans la codification du caractère arabe :

- ❑ Les caractères de base de l'alphabet arabe avec leurs différentes formes contextuelles ;
- ❑ Les caractères de voyellation ;
- ❑ Les caractères latins pour la représentation de l'arabe parlé ;
- ❑ Une amélioration des caractères de contrôle.

[64] L'Arab Technical Group est le groupe de l'ECMA qui a la charge de développer cette partie de la norme et de collaborer avec l'ISO (dans le cadre des travaux du groupe de travail JTC/SC2/WG2) afin de veiller à la prise en compte des commentaires de l'ECMA.

Le jeu des caractères UCS (ISO/IEC 10646) réserve ainsi une partie du BMP au codage des caractères arabes au niveau de la sixième rangée du BMP (en contiguë, les caractères arabes de base et étendus). Les rangées FB, FC, FD et une partie de la rangée FE sont destinées à coder les différentes formes de présentation des caractères arabes : isolés, surimprimés (overstiking), sur deux caractères ou plus (Voir [Annexe I](#)).

Parmi les particularités qui distinguent le caractère arabe dans le BMP on relèvera l'aspect de la surimpression des caractères composés et le phénomène de la bidirectionnalité.

3.4.2.1 Les caractères composés (Overstiking)

Quand un ordinateur traite des données textuelles, le traitement s'accomplit sur la plus petite unité d'information possible identifiée comme éléments de texte. Un élément de texte dépend de la nature de traitement qu'il subit. La combinaison arabe « Lam-Alef » est un seul élément de texte pour une opération de rendu (affichage & impression). Elle devient par contre deux éléments de texte « Lam » puis « Alef » pour d'autres opérations comme le tri.

Dans d'autres cas, la même opération réalisée sur la même chaîne de caractères nécessiterait des éléments de texte différents en fonction de la langue associée aux chaînes de caractères. C'est le cas par exemple de la chaîne « ch » qui varie en fonction de son appartenance à une langue ou à une autre. En français, le tri de la chaîne considérerait les caractères « C » et « H » séparément. En espagnol par contre, l'élément de texte trié est le caractère composé « CH » en tant que caractère unique.

Certains caractères codés par Unicode sont des caractères composés : c'est-à-dire qu'il s'agit de l'association stricte (affichés ou imprimés) de certains caractères (i.e. les accents en français ou les voyelles en arabe) aux caractères qui les précèdent dans le même espace. À titre d'exemple, le caractère ê (la lettre e avec un accent circonflexe dessus) est composé d'une forme de caractère « e » et d'une autre forme du caractère « ^ » non espacé ^[65] qui lui est superposé.

Un caractère composite est typiquement constitué d'un caractère de base qui occupe un seul espace est un ou plusieurs caractères non espacés qui prennent la même position de façon à ne pas créer un encombrement mais plutôt à compléter la forme du caractère de base dans une perspective graphique. Unicode spécifie l'ordre dans lequel ces caractères apparaissent de sorte que le caractère de base s'affiche en premier lieu suivi des caractères non espacés.

Dans le cas où un caractère composé nécessiterait plus d'un caractère non espacé, l'ordre de leur apparence ne constituera un problème que dans la mesure où ils occuperaient la même position. Unicode dispose dans ce cas de figure d'un mécanisme de priorité attribué à chaque caractère de base.

Unicode permet une deuxième possibilité pour gérer les caractères composés : les caractères précomposés. Chaque caractère précomposé (ou préfabriqué) dispose dans ce cas, d'une seule combinaison binaire (ou point de code) au lieu de deux ou trois qui constitueraient le caractère composé au moment de l'affichage ou de l'impression. Le cas du caractère ü (un « u » avec un tréma dessus) peut être codé avec un simple code U+00FC (=ü) ou avec le point de code de « u » suivi du point de code non espacé U+0308 du tréma. Unicode a adopté ce mécanisme de caractères précomposés pour garantir un maximum de compatibilité avec des standards

[65] Par caractère non espacé on sous-entend un caractère graphique qui, en s'inscrivant à partir du clavier, n'engendre pas un décalage de l'espace d'écriture vers l'emplacement suivant (gauche ou droite).

précédents comme ISO/IEC 8859-1 qui incluent beaucoup de caractères composés comme ü (u avec tréma), ñ (n avec tilde)... La langue arabe est riche en caractères composés. Unicode leur a attribué une zone particulière au niveau des rangées FB, FC, FD et FE du BMP (voir [Annexe I](#)).

3.4.2.2 La bidirectionnalité

La bidirectionnalité est l'une des difficultés majeures des systèmes multilingues. Alors que certaines écritures comme le latin ou le cyrillique sont écrites de gauche à droite, certaines autres écritures le sont de droite à gauche comme l'arabe et l'hébreu. La directionnalité n'est pas limitée à l'affichage des caractères (glyphes) des langues appropriées uniquement. Certains caractères communs de ponctuation nécessitent également un traitement particulier. C'est le cas des paires de caractères symétriques ([Tab.22](#)) qui, tout en gardant la même apparence, changent d'orientation avec le changement du sens de l'écriture. Un crochet d'ouverture en mode latin « [» deviendrait par le phénomène de pivotement (swapping) un crochet de fermeture en mode arabe et vice versa. Ceci est d'ordre à influencer également les opérations de tri et de recherche des données textuelles multilingues (§ [IV-2.2](#)).

Pour traiter cet aspect d'opposition de sens de l'écriture, Unicode dispose d'un algorithme BIDI qui attribue à chaque caractère une propriété directionnelle conformément à l'identité du script auquel il appartient. Les lettres latines sont d'une directionnalité gauche-droite ; les caractères arabes et hébreux sont d'une orientation droite-gauche.

Tous les caractères imprimés sont ainsi classés par Unicode selon la nature de leurs attributs de directionnalité :

- ❑ Gauche-droite forte (Mode Maître Latin) ;
- ❑ Gauche-Droite faible (Mode Insertion Latin) ;
- ❑ Droite-Gauche forte (Mode Maître Arabe) ;
- ❑ Droite-Gauche faible (Mode Insertion Arabe) ;
- ❑ Neutre.

En plus, Unicode définit le concept de la directionnalité globale associée à un bloc de texte qui est approximativement identique à un paragraphe. L'une des tâches primaires du programme d'affichage ou d'impression est de déterminer l'orientation globale du texte qui sera celle d'un mode maître d'édition. Les incises de chaînes de caractères dans d'autres écritures sont traitées conformément à leurs attributs de directionnalité. Les caractères neutres prennent les attributs des caractères qui les entourent et s'affiche dans une directionnalité relative.

Le traitement de la bidirectionnalité sous Unicode est automatiquement accompli par les algorithmes par défaut. Cependant, dans des situations de complexité avancée, des indications supplémentaires sont parfois nécessaires pour indiquer une directionnalité forcée. Pour cette raison, Unicode dispose d'un ensemble de règles et de caractères invisibles de codage (Code U+200E pour la direction Gauche-Droite et code U+200F pour la direction Droite-Gauche) que les logiciels d'affichage ou d'impression interprètent pour établir la direction de chaque séquence de caractères :

- ❑ Marque de Gauche à droite (Left-to-Right Mark : LRM) ;
- ❑ Marque de Droite à Gauche (Right-to-left Mark : RLM) ;
- ❑ Insertion de Droite à Gauche (Right-to-left embedding : RLE) ;
- ❑ Insertion de Gauche à Droite (Left-to-right embedding : LRE) ;

- ❑ Remplacement de Gauche à Droite (Left-to-right override : LRO) ;
- ❑ Remplacement de Droite à Gauche (Right-to-left override : RLO) ;
- ❑ Formatage directionnel POP (Pop directional formatting : PDF).

Le rôle des algorithmes d'affichage (rendering) est effectivement nécessaire pour traiter à la fois les questions de la bidirectionnalité et de l'analyse contextuelle pour la sélection des glyphes appropriés en fonction de la position du caractère dans le mot.

Unicode traite le changement de la direction des textes horizontaux uniquement. Le tableau suivant illustre les propriétés de la bidirectionnalité attribuées aux différentes écritures couvertes par Unicode.

Code	Direction	Famille de caractères	Rangées
L (Left)	Gauche-Droite	La plupart des caractères alphabétiques, syllabiques, idéographiques :	
		Lettres latines :	U+0041 → U+005A, U+0061 → U+007A, U+00C0 → U+00D6, U+00D8 → U+00F6, U+00F8 → U+00FF
		Latin Européen → Modificateurs de Lettres :	U+0100 → U+02FF
		Diacritiques généraux :	U+0300 → U+036F
		Grec → Arménien :	U+0370 → U+058F
		Devanagari → Géorgien	U+0900 → U+10FF
		Hiragana → Han :	U+3040 → U+8BFF
		Chiffres romains :	U+2160 → U+2182
		Marque Gauche-Droite :	U+200E
		Symboles diacritiques :	U+20D0 → U+20FF
		Divers :	U+0026 → +0040
R (Right)	Droite-Gauche	Écritures arabe, hébreu et ponctuations qui leur sont spécifiques	
		Arabe et hébreu :	U+0590 → U+065F, U+066D → U+06EF
		Marque Droite-Gauche	U+200F

Tableau 27 : Répartition des caractères Unicode selon la direction de l'écriture

Il est à noter que les codes de bidirectionnalité ne sont interprétés que dans le cas de textes bilingues horizontaux. Ils ne sont pas affichables et n'entrent pas donc dans les opérations de tri et de comparaison des chaînes des caractères.

Si les codes LRM et RLM spécifient les attributs d'une directionnalité forte (mode Maître), les codes à directionnalité faible (RLE, LRE, LRO, RLO) ou neutre, déterminent la direction de leurs rendus à partir des caractères qui les entourent.

D'autre part, si les codes RLM et LRM définissent la directionnalité des modes maîtres respectifs, les codes d'insertion RLE et LRE ainsi que les codes de remplacement LRO et RLO introduisent des sous chaînes de caractères de directionnalités opposées. Le sous chaînage des caractères à directionnalités variables peut aller dans certaines applications jusqu'à 15 niveaux hiérarchiques dans lesquels chaque code RLE, LRE, LRO, ou RLO introduit un niveau inférieur et chaque code PDF effectue un retour au niveau précédent. La directionnalité du niveau supérieur est implicite ou déterminée par l'application en cours.

Pour Unicode, l'ordre logique de l'enregistrement des caractères est celui de l'ordre de leur saisie sans aucune considération du sens de l'écriture de leurs langues d'origine.

Nous prendrons l'exemple d'un texte saisi dans une langue latine. Les caractères vont être saisis dans un enchaînement Gauche à Droite. Un texte en langue arabe serait par contre une suite de caractères enchaînés de Droite-Gauche. Un texte bilingue en mode maître arabe serait quant à lui une suite successive de caractères de Droite à Gauche avec des insertions occasionnées de caractères latins de Gauche à Droite. Dans les deux cas, Unicode ne verrait que l'ordre séquentiel dans lequel les caractères sont saisis.

4. LE CODAGE DES CARACTERES DANS LES SYSTEMES OUVERTS ET DISTRIBUES : LE CAS INTERNET

Nous avons abordé jusqu'à ce point les éléments du traitement multilingue d'un point de vue normatif et dans une abstraction totale de leurs domaines et formes d'application dans les solutions informatiques là où le conflit et la diversité linguistique prennent normalement forme.

S'il est vrai que la problématique des codes des caractères à été largement résolue au niveau des contextes de travail en mode monoposte ou affichant une uniformité matérielle (même plates-formes), applicative (même logiciels) ou culturelle et linguistique (même zone géographique ou même culture linguistique), la question du codage des caractères et des extensions des codes utilisées pour étendre les champs d'application des systèmes multilingues ne sauront à notre sens être complets si nous passons sous zone d'ombre le contexte Internet. L'importance de ce nouveau phénomène universel nous est évidente en raison des considérations suivantes :

- ❑ Internet devient le standard de fait de la majorité des contextes de travail des individus et des institutions, des profanes et des experts, du secteur privé et du domaine public, du monde de la recherche et de l'éducation ainsi que de l'univers des entreprises... bref de toutes entités ayant besoin d'utiliser ou de diffuser un produit d'information quelconque ;
- ❑ Internet est l'apanage des toutes les solutions informatiques du monde et des plates-formes de travail existantes depuis les machines portables jusqu'aux grands systèmes téléinformatiques ;
- ❑ Internet est le plus grand gisement d'information du monde qui couvre tous les secteurs possibles du savoir humain ;
- ❑ Internet est un lieu de rencontre de toutes les cultures et de toutes les langues du monde quoi que l'anglais y constitue la *lingua franca*;
- ❑ Internet est porteur d'une nouvelle technique de travail partagé, en l'occurrence l'architecture client-serveur. Ce modèle constitue actuellement le pilier fondamental des systèmes ouverts, répartis et distribués adoptés par tous les systèmes d'information et de communication ;
- ❑ Internet est le milieu favori des applications groupware et des documents composites ;
- ❑ Internet est le croisement des échanges et des transactions volumineuses d'information inter-individus, inter-groupes et inter-institutions.

Pour toutes ces raisons ainsi que tant d'autres, nous donnerons dans cette partie plus d'intérêt au phénomène Internet qu'aux solutions individuelles (i.e. Bureautique, Réseaux Locaux..) où le choix linguistique n'est pas confronté à des problèmes majeurs de diversité de langues et de virtualité d'utilisateurs. Les conditions de travail sur une application de Bureautique ou dans un environnement réseau local sont généralement identifiées, maîtrisées et paramétrées pour

réponde à des besoins limités et bien précis (voir en guise d'exemple le paramétrage multilingue des machines PC dans [l'Annexe D1](#)).

Internet est par contre le monde du virtuel et du concret, du cosmopolite et de l'unique, de l'invisible et du palpable, du conventionnel et de l'aléatoire, de l'international et du local. Bref c'est le melting-pot de la technologie et de l'information où la transparence tant au niveau des infrastructures que des superstructures constitue le soubassement fondamental. Le seul point d'achoppement qui reste à surmonter est la transparence linguistique. Dans un univers virtuel comme Internet comment pourrions nous prévoir le profil linguistique d'un fournisseur d'information, comment parviendrons-nous à analyser la requête d'un quémendeur de ressources, comment identifierons nous la structure documentique d'une donnée si les soubassements de ce monde labyrinthique est incontrôlable tout du moins sur cet aspect des langues d'interfaces ? La réponse à cette question passe obligatoirement par les phénomènes de l'internationalisation et de la localisation en pleine effervescence au niveau des services Internet.

Avant d'aborder ces deux aspects, exposons tout d'abord les aspects les plus significatifs pour notre étude du fonctionnement d'Internet. Bien que cela ressort du cadre de notre étude actuelle, nous en parlerons en termes de World Wide Web^[66], protocole désormais fédérateur de toutes les formes et les émulations des protocoles de navigation sur Internet comme le Wais (Wide Area Information System), Gopher, Archie, FTP (File Transfert Protocol), Telnet, Mail, News, etc. Notre souci est de préparer le contexte d'analyse des points linguistiques que nous irons ensuite détecter dans le capharnaüm des coulisses Internet. Nous étudierons les aspects linguistiques de ces services sous l'angle de la nature et de la profondeur des solutions multilingues apportées afin de garantir la transparence linguistique universelle qui cherche encore le statut d'un consensus général.

4.1 LE SCHEMA DE FONCTIONNEMENT DU WORLD WIDE WEB

Le World Wide Web canalise les informations échangées sur Internet (texte, image, vidéo, son) selon une répartition des tâches entre des machines dites serveurs et d'autre dites clientes. C'est la nouvelle architecture ouverte des systèmes distribués qui a remplacé l'ancienne architecture « terminal-serveur » des années 50 démunie de toute forme de partage des tâches et d'intelligence du côté de l'utilisateur. Le schéma global de travail du WWW est reparti entre deux catégories d'intervenants :

- ❑ un navigateur doté de toutes les potentialités de repérage (recherche), identification, filtrage, accès, négociation, récupération et présentation des ressources d'information,
- ❑ une série de protocoles de communication standardisés que W3 met en service pour gérer l'adressage, la négociation entre ressources, l'acheminement et le contrôle des transferts entre machines.

Plus précisément, ce schéma global de fonctionnement s'inspire du modèle client-serveur pour contrôler les communications sur le réseau Internet et partager les tâches et les étapes d'accès à l'information entre une machine cliente et une machine serveur selon les étapes suivantes :

- ❑ Le « client », la machine qui demande l'accès distant, envoie une requête à un « serveur » W3 pour demander la réception à travers le réseau d'un document quelconque (texte, image, séquence vidéo, son ou un document multimédia composite). Ceci s'effectue par

[66] Aussi identifié par WWW ou W3. En français l'expression « toile d'araignée mondiale » est utilisée bien que l'expression anglaise est plus généralisée.

la précision d'une adresse électronique du document ou par simple clic sur une zone de lien dynamique (hypertexte) qui envoie la commande de manière automatique.

- ❑ Le programme d'accès aux ressources WWW valide l'adresse spécifiée au niveau d'un serveur de noms (DNS=Domain Name Server) qui lui confirme la possibilité de router la requête vers le serveur W3 en question.
- ❑ Sur réception de la commande formulée par le client, le serveur envoie en retour les documents (fichiers texte et/ou images) au client sous contrôle d'un protocole particulier de transfert de documents hypertexte, en l'occurrence le protocole HTTP (HyperText Transfer Protocol).
- ❑ Une fois transmis, le document est mis en page puis restitué (visualisation, exécution dans le cas d'animation...) par le navigateur installé sur la machine du client. La charge de travail est ainsi répartie entre un client (quémandeur d'information) et un serveur (celui qui la fournit).

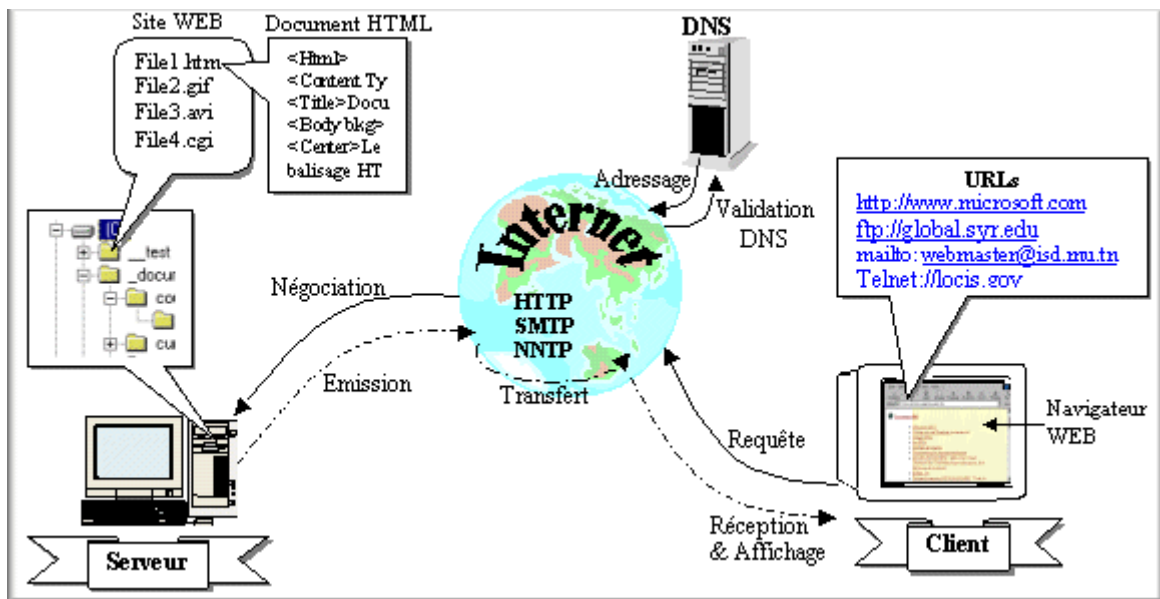


Figure 28 : Schéma de fonctionnement du WWW

Ce schéma, bien que relativement détaillé, n'est en fait que la partie visible de ... l'opération décrite. En réalité, d'autres protocoles, transparents pour l'utilisateur, interviennent à des niveaux différents du processus de communication pour gérer une série de tâches de fonds. Ces protocoles, définis entre autres par le NCSA^[67], le CERN^[68], l'IETF^[69], et devenus par la force d'usage des standards de fait, ont permis de mettre au point des procédures et des mécanismes uniformes d'échanges et de présentation des documents en hypertexte ainsi que l'adressage universel des documents et des services électroniques. Il s'agit :

- ❑ des adresses URL (Uniform Resource Locators) : ou la définition unique d'adresses permettant de localiser une ressource. Les URL peuvent être rattachées à plusieurs protocoles : « http » pour les documents HTML, « smtp » pour les services de

[67] National Center for Supercomputing Applications. Université de l'Illinois, Urbana-Champaign. <http://www.ncsa.uiuc.edu/General/NCSAHome.html>

[68] Centre Européen de Recherche Nucléaire à Genève. <http://www.cern.ch>

[69] L'Internet Engineering Task Force (<http://www.ietf.org/> voir aussi <http://www.ietf.cnri.reston.va.us>) est la partie chargée de l'ingénierie et du développement d'Internet. Il s'agit d'une communauté internationale ouverte de concepteurs de réseaux, d'opérateurs, de vendeurs et de chercheurs impliqués dans l'évolution de l'architecture Internet et de son fonctionnement.

messagerie, « nntp » pour les adresses des news, « ftp » pour le transfert de fichiers etc. dans une structure particulière ;

Fichier local	<code>file:///repertoire/fichier.txt</code>
Web	<code>http://serveur:port/repertoire/fichier.html</code>
Ftp	<code>ftp://serveur/repertoire/fichier</code>
Messagerie	<code>mailto:nom@organisation.domaine</code>
Telnet	<code>telnet://Nom:Password@serveur:port</code>
Gopher	<code>gopher://serveur:port/repertoire/fichier#marqueur</code>
Usenet	<code>news://serveur:port/repertoire/nom.de.la.news</code>
Bases Wais	<code>wais://<host>:<port>/<database></code>

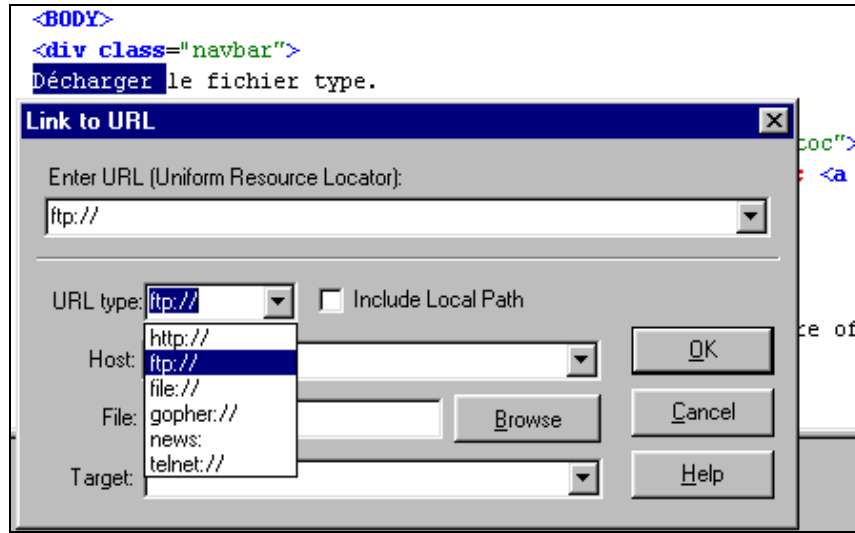


Figure 29 : Choix des protocoles sous l'éditeur Htmled32 de 2^e génération

- ❑ du protocole HTTP (HyperText Transfer Protocol) : Un protocole rapide et sûr, aux capacités étendues, permettant le transfert de documents multi-formats,
- ❑ du langage HTML (HyperText Markup Language) : une syntaxe spécifique à WWW qui permet de baliser un document pour lui donner une forme et une structure compréhensibles par tous les clients.

En réalité, tout Internet est géré à niveau de ces trois axes fondamentaux ; car en fait chacune des ces trois composantes représente un ensemble de protocoles uniformes qui interviennent à des strates différentes des processus d'usage et de traitement de l'information sur Internet :

- ❑ Les URIs (Universal Resource Identifiers = RFC 1630, RFC 2396)^[70] utilisés pour décrire les noms et les emplacements d'objets (§ III-4.2.1.2). Les URLs (RFC 1738^[71], RFC 1808)^[72] et les URNs (Uniform Resource Numbers = RFC 1738, 1808)^[73] en sont des exemples. Les URIs, également identifiées comme UDI (Universal Document Identifiers) ou tout simplement comme une combinaison entre les URLs et les URNs prennent forme d'une chaîne de caractères disposés selon une structure particulière fixée

[70] « Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web », RFC 1630, CERN, Juin 1994.
http://www.w3.org/pub/WWW/Addressing/URL/URI_Overview.html

[71] RFC 1738 = <ftp://ds.internic.net/rfc/rfc1738.txt>

[72] Voir T. Berners-Lee, L. Masinter, M. McCahill. Uniform Resource Locators (URL), December 1994. URL : <http://www.w3.org/pub/WWW/Addressing/URL/Overview.html>

[73] Ou aussi « Uniform Resource Names ».

par le RFC 1630^[74] pour identifier une ressource quelconque à travers un nom, un emplacement ou toute autre caractéristique propre au document concerné. Les URCs (Uniform Resource Citations^[75]) établissent une jonction entre les URNs et les URLs à travers une série de propriétés (métadonnées) en forme de champs uniformes de description du document (titre, auteur, date, éditeurs...). Certaines valeurs d'URC peuvent être des URIs. Le changement des URLs peut ainsi être suivi par des usagers autorisés (i.e. webmasters) au sein même du service URC.

- ❑ Le protocole HTTP est un porte-étendard d'une catégorie particulière de protocoles Internet qui cogitent dans les méandres du réseau des réseaux à l'insu de l'utilisateur final. Si HTTP, comme son nom l'indique, gère le transfert et la communication des documents en hypertexte, SMTP (Simple Mail Transfer Protocol) est le garant des transactions entre plates-formes hétérogènes de messagerie électronique. Le protocole NNTP (Network News Transfer Protocol), défini par RFC 977, est un protocole de distribution, de recherche et de récupération d'articles de « news » entre une communauté d'adhérents aux listes des Nouvelles éparpillées sur Internet.
- ❑ Le langage HTML est un dérivé de SGML (Standard Generalized Markup Language) défini par la norme ISO 8879 (1986). Par opposition à HTML qui consiste à marquer un texte pour contrôler son formatage visuel et imprimé, SGML consiste à définir la structure interne d'un document par la description logique des différentes parties qui le composent. SGML a évolué vers une nouvelle norme de description logique, en l'occurrence le XML (Extensible Markup Language) dont l'objectif global est de faciliter davantage l'installation, l'usage et le traitement des documents SGML à la manière dont il est couramment fait de HTML. Pour cette raison, XML est conçu sur des principes d'interopérabilité avec HTML et SGML à la fois.

Pour cette raison, nous concentrerons notre intérêt sur ces trois phases principales du mécanisme Internet pour détecter toute forme de norme ou standard de traitement multilingue.

4.2 LA DIMENSION MULTILINGUE DANS LES PROTOCOLES INTERNET

Le mécanisme fonctionnel d'Internet, comme nous l'avons précisé dans la partie précédente, est axé sur les trois pôles d'adressage, de routage et de traitement.

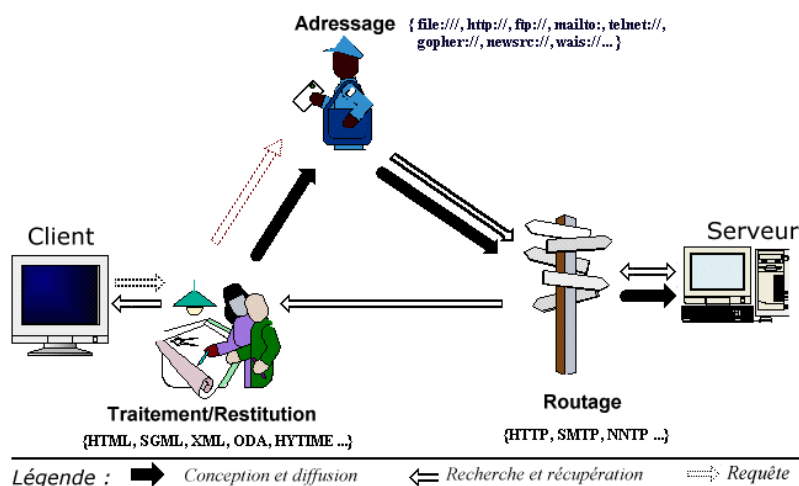


Figure 30 : Articulation entre protocoles Internet

[74] T. BERNERS-LEE. « Universal Resource Identifiers in WWW ». Op. cit.

[75] Ou également « Uniform Resource Characteristics ».

Au niveau de l'un ou de l'autre de ces trois strates du mécanisme Internet, le volet linguistique a toujours été d'un penchant latin et plus particulièrement anglo-saxon. Avec la généralisation du modèle client–serveur et l'éclatement des barrières géographiques des services Internet pour donner lieu à des concepts comme le « village planétaire », la mondialisation et la globalisation des systèmes d'information, une présence linguistique hétérogène est devenue indispensable pour permettre la présence de toutes les cultures et les langues du monde dans cette réalisation globale et globalisante qu'est le réseau Internet. Les premières réactions provenaient du monde des langues utilisant le caractère latin (i.e. français, allemand ...) pour combler le vide enregistré au niveau du traitement des caractères accentués (voir parties précédentes dans ce chapitre pour les détails concernant les extensions de codes). Aujourd'hui, la langue française ou toute autre langue à diacritiques est techniquement équivalente à la langue anglaise. Le seul handicap qui les distingue est la primauté historique que l'anglais a gagnée dans le développement des logiciels et la conception des contenus d'information.

Or si l'on parle d'internationalisation et de localisation, il est évident que la liste linguistique ne devrait pas s'arrêter aux langues latines et germaniques exclusivement. Deux questions fondamentales s'imposent à ce niveau :

1. Quelle est la marge de présence des autres langues dans cet univers virtuel ?
2. À quel niveau les solutions de l'internationalisation et de la localisation courantes assurent-elles la présence des langues dites « vernaculaires » ?

De prime abord, l'aspect vers lequel tendrait toute tentative de valorisation des solutions d'internationalisation est sans doute le volet contenu. En d'autre terme, si l'on parle de multilinguisme d'Internet, il est couramment sous-entendu que l'on évoque l'aspect linguistique de son contenu informationnel plutôt que de ses outils de fonctionnement.

Or l'un des problèmes encore sans solution évidente est le problème de l'adressage. Toute ressource sur Internet, située sur un nœud accessible à distance est identifiée par une adresse unique (URL=Uniform Resource Locator) qui sert d'une part de pointeur pour l'utilisateur final, et d'autre part de point d'entrée dans les index et les répertoires des moteurs de recherche ; donc qui permet sa récupération par des outils de recherche d'information. La situation actuelle est encore très restrictive et exclusivement limitée aux caractères latins et plus particulièrement les caractères ASCII. La question est de savoir où en est la situation avec des langues autre que les latines et plus particulièrement la langue arabe.

Nous aborderons dans les parties qui suivent l'internationalisation (d'un point de vue linguistique) des ressources et services Internet au niveau de trois paliers qui le constituent : l'adressage, le routage et la structuration des documents électroniques conçus et échangés sur le réseau des réseaux.

4.2.1 L'Internationalisation des URL

On appelle URL une adresse généralement mnémonique qui identifie une ressource accessible par Internet. La structure générale d'une adresse URL se divise en trois parties distinctes :

- 1) un protocole spécifique à la nature de la ressource à récupérer (http, gopher, ftp, mailto, news, telnet etc.);

- 2) un nom de domaine (nom du serveur) qui héberge la ressource concernée. Il est parfois identifié par son adresse numérique IP (Internet Protocol) attribuée par le serveur de noms de domaines (DNS = RFC1034^[76], RFC1035^[77]);
- 3) un chemin d'accès qui permet de localiser la ressource sur le serveur.



Figure 31 : Structure d'une adresse URL

L'adresse URL, comme la démontre sa structure illustrée dans la figure précédente, joue en fait le rôle d'un nom d'une ressource informatique (i.e. fichier) accessible sur une machine distante ou locale (domaine) à travers son chemin arborescent grâce à un protocole approprié. La dominance historique des caractères ASCII bien qu'elle ait évité l'anarchie dans la détermination des noms des domaines (DNS = RFC 1034 & RFC 1035) et la définition des URLs (RFC 1738), a engendré en contre partie des limites dans le choix des caractères utilisés pour les dits paramètres. Les tentatives pour remédier à cette limitation sont connues sous le nom d'Internationalisation (i18n^[78]).

L'internationalisation des URLs, bien que latente pour la grande majorité des utilisateurs d'Internet, constitue le plus grand problème actuellement rencontré par les divers opérateurs techniques d'Internet. Les questions fondamentales encore abordées dans les documents techniques de l'IETF (Internet Engineering Task Force) et des groupes de travail qui en dépendent (i.e. IRTF^[79], IAB^[80], ISOC^[81], IANA^[82]), concernent les règles de codages des caractères (ou des octets) à mettre dans la zone des URLs et les conventions utilisées pour les représenter. Aucun consensus général n'est encore atteint sur cet aspect. Faisons en un topo critique. (N.B. : Nous nous limiterons au protocole HTTP comme échantillons de l'i18n des Urls. Pour voir les spécificités du protocole FTP, nous recommandons de voir l'avant projet (Internet Draft) de B. Curtin^[83]).

[76] P. MOCKAPETRIS. Domain names – Concepts and facilities. Network Working Group, RFC 1034. November 1987.

[77] P. MOCKAPETRIS. Domain names – Implementation and specification. Network Working Group, RFC 1035. November 1987.

[78] i18n : abréviation du concept « Internationalisation » qui consiste à insérer entre l'initiale « i » du terme et sa terminaison « n » le nombre « 18 » des caractères qui les séparent. L'abréviation est très utilisée sans le contexte Internet.

[79] L'Internet Research Task Force (<http://www.isi.edu/irtf/>) est composé de groupes de recherche œuvrant sur des sujets relatifs aux protocoles, applications, architecture et technologies d'Internet.

[80] L'Internet Architecture Board (<http://www.iab.org/iab/>) est membre de l'Internet Society, responsable des considérations générales de l'architecture d'Internet.

[81] L'Internet Society (<http://info.isoc.org/>) est un Organisme Non Gouvernemental International pour la coopération et la coordination dans Internet.

[82] L'Internet Assigned Numbers Authority (<http://www.iana.org/iana/>) est le coordinateur central de l'assignation des paramètres uniques pour les protocoles Internet.

[83] B. CURTIN. Internationalization of the File Transfer Protocol. FTPEXT Working Group. Internet Draft, juin 1998. <http://...draft-ietf-ftpext-intl-ftp-05.txt>

4.2.1.1 État de la question

Les restrictions linguistiques de la définition des adresses URL rappelle encore le contexte de la nomenclature des fichiers sous MS-DOS dans le sens où seuls les caractères ASCII peuvent être utilisés. En effet, les URLs sont généralement limitées à l'usage d'un sous-ensemble de caractères du répertoire US-ASCII (de A à Z casses confondues, de 0 à 9 et le tiret). Le reste des caractères ASCII ainsi que des caractères non-ASCII sont reproduits à travers les techniques de codage particulier (voir format %HH dans § [III.4.2.1.2.3](#)). Le problème des caractères non ASCII réside dans la non identification par les serveurs des jeux des caractères auxquels ils appartiennent.

Les URLs, selon la définition du RFC 1738, sont des séquences de caractères (lettres, chiffres et caractères spéciaux) utilisées pour représenter des séquences d'octets reconnues par les protocoles Internet. Certaines de ces séquences d'octets sont représentées à travers un mécanisme de codification particulier pour représenter des caractères non-ASCII. En définitive, la définition d'une URL peut s'obtenir à travers une ou les deux méthodes suivantes à la fois :

- ❑ Une représentation directe des octets par leurs équivalents en caractères ;
- ❑ Une représentation indirecte des octets par leurs équivalents codés en séquences d'échappement ;

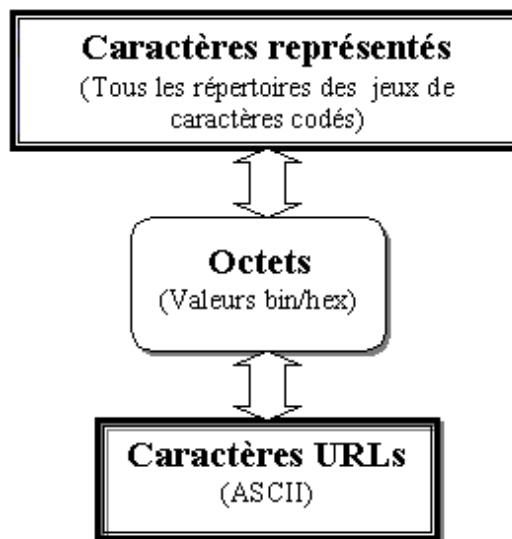


Figure 32 : Niveaux de définition des caractères URL

Le statut qui est encore en vigueur, est que la définition des URLs s'effectue exclusivement à base de caractères graphiques du jeu des caractères codés de l'US-ASCII. Ceci est de nature à défavoriser toutes les langues du monde, excepté celles transcrites en ASCII (l'Anglais et le Swahili), dans la transcription des adresses mnémoniques compréhensibles de leurs URLs. L'ambiguïté est que tout utilisateur préfère traiter directement avec des URLs mnémoniques qu'il pourrait imprimer, saisir sur machine ou transmettre dans le texte d'un message ou d'un document quelconque. Ceci va à l'encontre du mécanisme interne des systèmes automatisés qui préfèrent les valeurs numériques et codées afin d'éviter toute forme d'ambiguïté.

Certaines solutions à ce problème commencent à prendre forme comme résultats de quelques efforts épars sans toutefois aboutir à l'état de consensus général.

Parmi les solutions encore à l'état de recommandation ou complètement obsolètes :

- ❑ Certains proposent un accord universel pour une conversion unique entre caractères et octets. Cette proposition limiterait l'usage de caractères URLs, au répertoire ASCII et exclurait tout autre jeu de caractères étendus.
- ❑ Certains proposent d'utiliser des jeux de caractères propres avec toutefois l'adjonction à toute URL d'une information sur le jeu de caractères utilisé. Seulement ceci constituerait pour certains une surcharge de la zone d'adresse qui devrait rester très légère pour accélérer son traitement et par conséquent l'accès à la ressource demandée. Un autre inconvénient attribué à cette solution est l'absence de l'effet de compatibilité rétrospective.
- ❑ Une toute nouvelle proposition émerge sous forme d'une recommandation pour l'acceptation universelle d'un encodage unique conforme à la technique de transformation des caractères Unicode vers UTF-8 (voir § III-3.4.1.3.2). L'avantage majeur à cette proposition est qu'un consensus universel éliminerait la nécessité d'adjoindre aux URLs multilingues l'indication des jeux des caractères utilisés. L'usage d'Unicode impliquerait que toutes les langues pourraient être présentes dans la définition des adresses URLs. Les deux points négatifs à cette proposition sont d'une part la restriction encore très lourde aux caractères ASCII dans l'attribution des noms des domaines (partie importante du schéma des URLs) au niveau des serveurs DNS, et d'autre part le problème de la compatibilité rétrospective des URLs non-ASCII et non-Unicode.

L'objectif de ces tentatives de solutions converge tout du moins vers l'extension des mécanismes courants pour la définition des adresses URLs dans le but d'identifier les données auxquelles elles pointent sans être obligées de s'appliquer ni aux caractéristiques linguistiques des machines sur lesquelles elles sont hébergées ni à la restriction aux caractères ASCII pour identifier un serveur. Nous reprenons en ce qui suit quelques-unes des propositions de l'internationalisation des URLs recensées dans la littérature spécialisée.

4.2.1.2 L'i18n des URLs : la verticalité de l'ASCII Vs l'horizontalité d'Unicode

« ... we ought to use a worldwide character set where even Japanese characters are integrity-preserved every where! Any language can be used internationally, not only English. We should not detour the issue. The net result to use ASCII worldwide plus a local character set only on a cultural island would be that English would dominate everywhere and other languages at a major disadvantage. This is not what I call internationalization, that is insularization »^[84]. Ainsi s'exprimait un intervenant dans un forum dédié à l'i18n des URLs pour mettre en exergue non pas uniquement la dominance anglo-saxonne dans le domaine de la définition des URLs, mais aussi pour dénoter le penchant toujours latin (restrictif) des solutions apportées à l'i18n de ce facteur.

Tout un débat est actuellement en cours pour justifier et argumenter des propositions et des modèles d'i18n des URLs. Certains sont basés sur l'équité de toutes les langues du monde pour favoriser leur présence au sein des URLs (solutions horizontales), d'autres voient la limitation à un seul jeu de caractère, en l'occurrence l'ASCII, comme solution à un allègement des URLs au moment de l'accès aux documents et comme garantie de compatibilité rétrospective avec les ressources déjà en place (solutions verticales).

[84] Alain LABONTE. Intervention au Forum « URL Internationalization » le mercredi 19 février 1997 09:37:07-0500. <http://www.acl.lanl.gov/URI/archive/uri-96.messages/>

Une première ébauche de solution pour l'i18n des URLs était proposée par le CW3 (Consortium WWW)^[85] qui recommande de manipuler les adresses des documents sur Internet par le truchement des trois « protocoles » URLs, URNs et URIs.

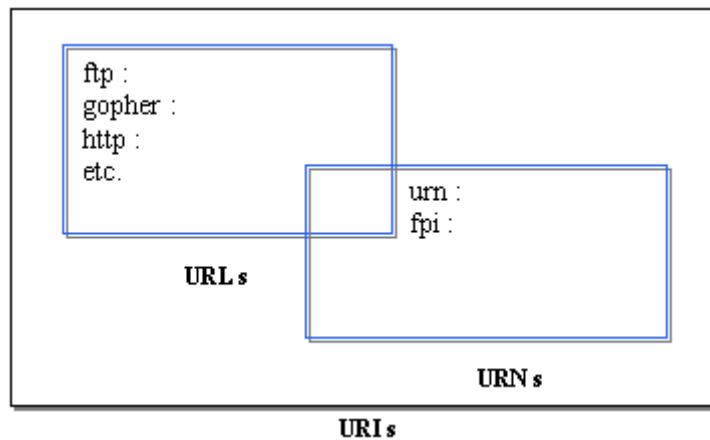


Figure 33 : Rapports URLs, URNs & URIs

Source : CW3. Naming and Addressing : URLs. <http://www.w3.org/pub/WWW/Addressing/Addressing.html>

Puisque les ressources sur Internet sont identifiées par des URIs, à leur tour composés par les URLs (les plus connus parmi les URIs) et les URNs (moins familiers), un système d'adressage multilingue pourrait être envisagé par la complémentarité entre les spécificités de chacun de ces deux derniers protocoles. Les URLs qui sont en fait une forme d'URIs dont la tâche est d'identifier la localisation d'une ressource quelconque sur le réseau peuvent compléter une URN dont la fonction principale est par contre d'identifier la ressource par son contenu global et unique. Les URNs (aussi identifiées comme Uniform Resources Numbers) peuvent identifier les données sous forme d'une chaîne de chiffres unique et univoque et surtout linguistiquement neutre. Ainsi l'URN servirait à identifier « *une ressource ou une unité d'information indépendamment de sa location, jouant en quelque sorte le même rôle que l'ISBN pour les monographies traditionnelles* »^[86].

Cette forme de complémentarité entre URLs et URNs nécessiterait de mettre en place un mécanisme de validation de la correspondance entre un identifiant « géographique » (URLs) et un identifiant sémantique (URNs). Les URNs numériques peuvent ainsi engendrer une meilleure persistance grâce à l'abstraction sémantique des chiffres avec toutefois le risque de confusion dans la correspondance entre URLs et URNs au moment de la conversion dans un sens ou dans un autre.

Dans ce cas de figure les DNS (Domain Name Servers = serveur de noms de domaines) pourraient servir de plate-forme intermédiaire. Le DNS est devenu un élément central de l'Internet pour gérer le mécanisme distribué de la vérification des paires nom-adresse des systèmes interconnectés et de leur attribuer des équivalents mnémoniques (nom de domaine : i.e. www.cck.rnu.tn) et numériques (Adresse IP : i.e. 193.95.66.10). L'objectif essentiel d'un DNS, comme défini par le RFC 1035, est de fournir un mécanisme de nommage de ressources capable de les identifier auprès de tous les serveurs, sur tous les réseaux et les protocoles Internet.

[85] Martin BRYAN. Web Internationalization & Multilingualism Symposium. Séville : WC3, 20-22 Novembre 1996. <http://www.sil.org/sgml/seville-w3c-int.html>

[86] Voir <http://www.enssib.fr/Enssib/resdoc/DOCUMENTS/DCB/samain/>

L'internationalisation des noms des domaines, souvent déclarée impossible, est hypothétiquement et techniquement rendue possible par Martin Dürst dans sa proposition (DRAFT) sous forme d'un RFC (<http://www.w3.org/international/draft-duerst-dns-i18n-00.txt>) et d'un projet pilote d'i18n d'URLs^[87].

Notre propre apport à cette méthode de jonction entre URLs et URNs, qui n'est d'ailleurs pas encore entrée en vigueur par manque de consensus de la part des constructeurs de matériel et des gestionnaires de réseaux, consiste à avancer une proposition que nous avons construite par la synthèse des contributions et des analyses éparses qui se construisent progressivement dans cette optique pour la mettre plus en harmonie avec la langue arabe au niveau des URIs et des noms des domaines.

D'emblée, nous considérons que l'importance du DNS dans sa forme numérique réside dans le fait de constituer un repère de validité qui a prouvé son efficacité et sa transparence linguistique depuis longue date. Il forme d'ailleurs l'une des parties fondamentales dans le fonctionnement et la structure des différents schémas d'adresses Internet.

4.2.1.2.1 Le modèle standard du fonctionnement DNS

Comme il est spécifié dans la RFC 1035, la plus simple configuration voire même la plus courante d'un service de noms de domaines est celle reproduite dans la figure suivante.

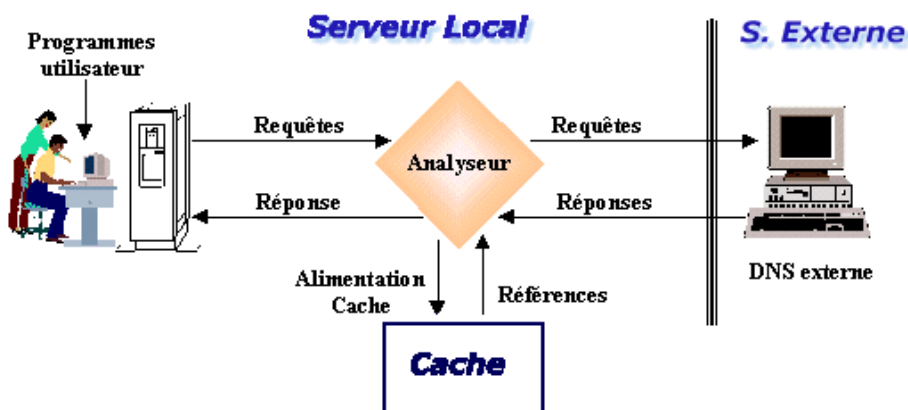


Figure 34 : Fonctionnement typique d'un service de noms de domaines

Adapté de : P. Mockapetris. Domain names – Implementation and specification. Network Working Group, RFC 1035. Novembre 1987

Les programmes utilisateurs dialoguent en réalité avec les DNS à travers un programme d'analyse d'adresses URLs (Resolver) conformément aux spécificités du serveur local qui formate les requêtes et les réponses en fonction de son système d'exploitation.

L'analyseur de requête construit sa réponse après consultation des serveurs de noms externes et de son système cache interne.

Dans un système où le serveur local accepte toutes les formes des flux des données supportées par un système DNS, l'interaction serait à double sens entre les requêtes en extrant et les requêtes en intrant (mode client et mode serveur).

[87] Voir la concrétisation du projet « IDNS » par le « Centre for Internet Research » (<http://www.cir.nus.edu.sg/>) à l'URL : <http://www.apng.org/idns/>

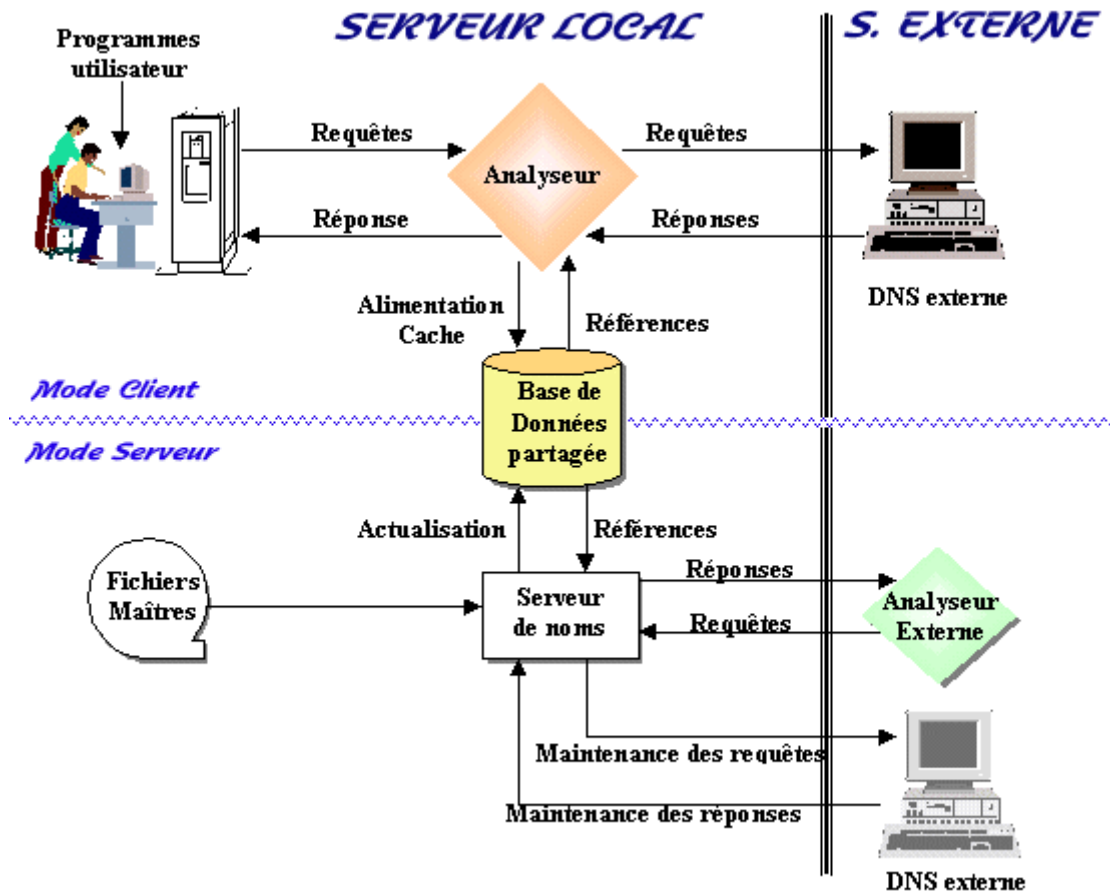


Figure 35 : Installation et fonctionnement standards des DNS

Adapté de : P. Mockapetris. Domain names – Implementation and specification. Network Working Group, RFC 1035. Novembre 1987

Le serveur de nom sur la machine locale pourrait être dans ce cas de figure un programme indépendant sur une machine réservée ou une procédure simple ou multiple installée sur le serveur central. Une base de données partagée entre le programme d'analyse des requêtes (resolver) et le serveur de noms, jouerait à la fois le rôle d'un système cache interne (données des requêtes précédentes) et d'un dépôt des données d'actualisation que le serveur de noms local effectuerait périodiquement à travers des circuits virtuels pour la mise à jour de ses registres. Pour répondre à une requête en intrant, le serveur de nom reconstitue et valide certaines zones d'un nom de domaine à partir de se son propre système de fichiers qu'il soumet en guise de réponse à une requête d'un analyseur externe. Celui-ci continuerait la validation de sa requête au niveau d'autres serveurs de noms externes. La requête d'un utilisateur quelconque est finalement traitée par l'analyseur de requête (resolver) au niveau de plusieurs serveurs de noms à la fois et impliquerait de la sorte plusieurs accès réseaux.

4.2.1.2.2 Notre proposition

Notre proposition partirait de cette configuration de base pour proposer l'adjonction de certains modules orientés vers le traitement linguistique des opérations standards des serveurs de noms et des analyseurs de requêtes. Nous rappelons que les limites ASCII dans la définition des noms des domaines était une contrainte originaire de l'exclusivité ASCII des systèmes d'exploitation d'antan surtout le système Unix qui constitue la plate-forme principale des serveurs de noms de

domaines actuels grâce à son programme BIND^[88]. Cette limitation était (elle l'est encore) d'ordre à empêcher l'usage de noms conviviaux même dans un contexte de travail multilingue. Or l'internationalisation des systèmes d'exploitation (y compris leur arabisation) a gagné du terrain pour couvrir la majorité des systèmes en usage actuellement (DOS, W95, Unix...). Nous croyons que si les serveurs de noms sont restés fixés sur les limites ASCII, leur internationalisation n'est pourtant pas de l'ordre de l'impossible. C'est une question de temps et surtout de contributions horizontales et équilibrées entre tous les partenaires linguistiques de la communauté Internet qui apporteraient progressivement les solutions adéquates. Notre proposition s'encadre dans cette perspective de proposer ne serait-ce que la filigrane d'une solution transparente future. Beaucoup d'expérimentations, d'investissements logistiques sont nécessaires pour mener des expériences de ce genre dans des laboratoires spécialisés et bien équipés.

a) Les principes de fonctionnement du modèle

Le fonctionnement de notre proposition se base essentiellement sur un mécanisme d'analogie, de conversion, de traduction et de validation des éléments constituant une adresse URL. Tout le mécanisme tourne autour d'une conversion unique et uniforme de toutes les zones du schéma d'une URL en valeurs numériques traitées actuellement de manière uniforme par tous les jeux des caractères codés. Notre inspiration, comme nous l'avons mentionnée auparavant émane du modèle du contrôle bibliographique universel à travers les valeurs ISBN (International Standards Book Number). C'est d'ailleurs la technique utilisée dans tous les systèmes de nommage par identification numérique (codes à barres, comptes bancaires, adresses IP ...). Cependant, à chaque bout du processus communicationnel, l'interface homme machine reste conforme aux exigences linguistiques de l'utilisateur qui pourra l'adapter à tout profil linguistique voulu. C'est entre les deux bouts de ce processus que tout se trame entre serveurs de noms, solveurs ou analyseurs de requêtes (resolvers), convertisseurs de codes de caractères, tables de correspondances URLs/URNs ... pour convertir, traduire et valider toutes sortes d'information transitant sur les réseaux.

La table de correspondance est la pierre angulaire de notre proposition car elle remplace à plusieurs niveaux le DNS conventionnel. C'est au niveau de la table des correspondances que s'effectueraient les traductions des URLs des utilisateurs vers leurs équivalents numériques.

Une adresse URL, comme illustrée auparavant, se compose de trois zones élémentaires : l'identifiant du protocole (http, ftp, mailto, news...), d'un nom de domaine et d'un identifiant de ressource. La phase de traduction s'effectuerait de différentes manières selon la zone à traduire.

1. **La zone protocole** : Étant donné que le nombre de protocoles est limité, celle-ci sera automatiquement traduite en valeur numérique par une routine greffée à la table de correspondance. Cette routine analyserait la première zone d'une URL transmise en codes de caractères unifiés (i.e. Unicode) et commune à toutes les tables des correspondances des systèmes interconnectés.
2. **La zone nom de domaine** : Elle est en stricte relation avec le serveur de noms (DNS) rattaché au système où la table des correspondances irait valider le choix de la valeur numérique d'un nom de domaine à inclure dans l'URL en phase de traduction (dans un sens ou dans un autre).

[88] BIND est le logiciel de références des serveurs DNS mis au point par la société Vixie Entreprises pour fonctionner en priorité sur les machines Unix. Il a été cependant porté sur les machines tournant sous Windows NT et OS/2.

3. **La zone de la ressource** : Elle est traduite dans la valeur numérique que le concepteur ou le détenteur lui aurait attribuée au moment de sa création ou de sa mise en service. Cette valeur est tributaire du nom du domaine auquel elle est rattachée.

En amont de cette phase de traduction, la partie mnémonique de la table de correspondance (URLs) est constituée par un programme de conversion des Codes de Caractères Particuliers ou personnels (CCP) vers un seul Code de Caractères Unifiés ou universels (CCU). Ce mécanisme fonctionne en double sens que ce soit dans un processus d'alimentation de la table de correspondance par des nouvelles entrées ou au moment de sa consultation pour analyse et validation d'adresses.

En aval de l'opération, la partie numérique (URNs) subit une traduction similaire mais en sens opposé pour lui redonner sa valeur mnémonique dans les jeux de caractères d'utilisateurs virtuels dans d'autres systèmes externes.

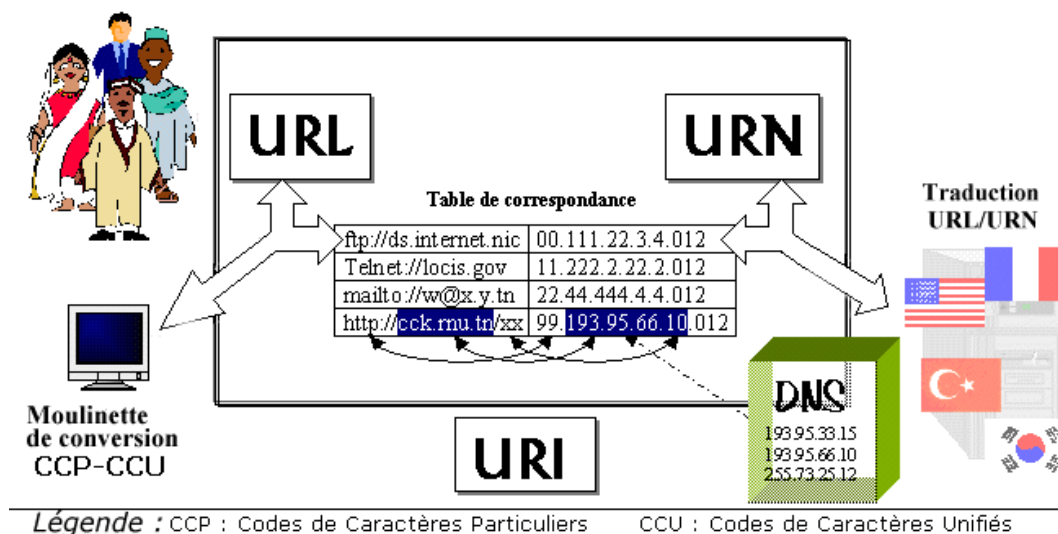


Figure 36 : Correspondance entre URLs et URNs via DNS

Ainsi conçu, tout utilisateur ayant accès à une interface Internet, dispose d'une plate-forme de travail qui correspond à son mode linguistique propre et, en mode optionnel, à un mode linguistique universel (i.e. ASCII ou Unicode). Ses requêtes en extrant (pointeurs URLs ou formulaires de recherche) ou les requêtes en intrant (idem de la part d'un utilisateur externe) subissent quand nécessaire (ce n'est pas le cas d'un CCP et d'un CCU identiques), en amont ou en aval du circuit communicationnel les mécanismes cités auparavant (la traduction URL/URN est toujours indispensable).

b) Le fonctionnement multilingue du mode client

Il reste à signaler que la définition et l'usage des URIs font face à deux situations différentes dépendantes de deux types d'identifiants (adresses) : un type d'identifiants de références à des données statiques et un type d'identifiants de références à des sources de données dynamiques.

Les données statiques sont susceptibles d'être consultées en permanence par l'utilisateur. Il est donc nécessaire qu'elles soient référencées de façon claire, lisible et facilement transcrite par tout lecteur potentiel. Ceci a toujours renforcé en quelque sorte l'exclusivité de l'ASCII pour les noms des domaines et pour les URLs partant du fait que l'ASCII constitue la base de tous les

jeux des caractères codés sur 8, 16 ou 32 bits. Mais, une critique d'usage est immédiatement formulée à l'égard de cette tendance de « sacraliser » cette limitation à l'ASCII dans les URLs : on peut se demander à titre d'exemple, à quoi servirait le fait qu'un utilisateur qui ne maîtrise pas la langue arabe puisse, comme dans l'exemple suivant, saisir en ASCII une adresse URLs d'un document arabe codé en code page arabe Windows 1256 ? L'inverse est aussi vrai.



Figure 37 : URL statique en ASCII pour un document en CP arabe Windows 1256

Dans un autre cas de figure, comment résoudre la problématique d'un utilisateur arabe unilingue qui aurait besoin de saisir une adresse URL d'un document arabe sur Internet ? Comment résoudre également la situation d'un utilisateur arabe ou chinois qui a besoin de saisir une adresse URL en ASCII d'un document dans sa propre langue en utilisant une plate-forme de saisie (i.e. clavier) unilingue ?

Bien que les deux situations soient techniquement solvables, car on peut toujours télécharger sur une machine connectée (ou avoir accès en ligne), une ressource en Java adaptée à un besoin linguistique (fonte, gestionnaire de claviers multilingues ...), le handicap peut parfois être d'ordre alphabétique. En d'autres termes, beaucoup d'individus dans le monde ne maîtrisent pas la forme des glyphes des caractères latins dans une URL qu'elle soit d'origine latine ou même locale translittérée.

La situation est certes très improbable étant donné le rapport souvent intrinsèque entre l'initiation à l'usage de l'outil informatique (vu son origine occidentale et latine) et l'apprentissage des caractères latins. Mais dans des contextes d'usage très spécifiques, un usager d'une communauté linguistique non latine, peut toujours exprimer un besoin d'aller chercher des données sur un navigateur dans une version linguistique qui lui est propre. Bien que minime, ceci est l'un des points encore à résoudre pour une ergonomie complète des interfaces multilingues Homme-Machine.

Sur un plan pratique cette approche, pour laquelle il faudrait un contexte de conception et d'évaluation assez lourd et complexe, présenterait plusieurs scénarios d'usage en fonction des deux situations suivantes :

1. Être générateur actif d'URL;
2. Être générateur passif d'URL;

En effet, dans un mode client, on peut être un générateur d'URL actif ou passif. Dans le mode actif, l'utilisateur est conscient de la ressource vers laquelle il pointe. Il en dispose d'une URL qu'il préciserait à un navigateur ou à un service d'accès quelconque (email, new, wais ...). Dans le mode passif, par contre, l'utilisateur n'est pas informé au préalable de l'URL à laquelle il aura accès. C'est le cas des situations d'envoi de formulaires de recherche qui génèrent des URLs spécifiques adressées aux moteurs de recherche et aux programmes CGI^[89] pour compilation interne. Les scénarios qui s'offrent à un client arabe dans ces deux cas de figure sont :

① **Scénario 1 : Travailler en mode client actif pour accéder à une ressource arabisée via une URL exclusivement arabe.**

Dans ce cas de figure qui correspond à un état de monolinguisme d'un utilisateur arabophone, la ressource arabe à atteindre est identifiée par une URL arabisée au niveau du client selon un codage de caractères particulier (CCP = ISO 8859-6, CP 1256). L'URL sollicitée par l'utilisateur est exprimée dans le CCP, traduite au niveau de la table de correspondance en URL mnémonique en CCU (i.e. Unicode) puis convertie en URL numérique. La table de correspondance contient déjà une entrée de la valeur numérique du serveur (DNS) avec les adjonctions numériques du protocole d'accès et du document référencé. La réponse (envoi de la ressource au client) s'effectue dans un sens inverse selon le même principe de codage, traduction et conversion entre les jeux des caractères codés. Bien qu'il s'agisse du même support linguistique (arabe), les codes de caractères entre le client et le serveur peuvent être différents. La phase de conversion reste donc toujours nécessaire sauf exception quand le CCP et le CCU sont identiques.

Cette opération reprendrait en quelque sorte le mécanisme actuellement en vigueur dans le contexte des formats d'échange bibliographiques qui s'accomplit à travers un Format Commun de Communication (CCF) pour échanger des notices bibliographiques entre systèmes hétérogènes.

Cet état de fait nécessiterait un pré-requis assez complexe pour le moment :

- ❑ Des noms de domaines en langue arabe. L'exclusivité actuelle des DNS aux caractères ASCII constitue un obstacle majeur même pour le passage aux caractères Latin-1 ;
- ❑ Des équivalents arabes à la partie protocoles d'accès dans le schéma des URLs actuels. Les expressions « http:// » ou « ftp:// » sont exclusivement en ASCII. Leur traduction en arabe implique un effort et un investissement considérables à part l'effet rétrospectif non garanti pour les ressources existantes. La traduction de cette partie de l'URL éviterait l'usage des balises « tags » de bidirectionnalité dans le cas d'URLs en écriture droite-gauche ;
- ❑ Un étiquetage des URLs pour identifier les jeux des caractères codés utilisés afin qu'ils soient convertis convenablement. Ceci constituerait une surcharge de l'élément URL par un balisage supplémentaire qui pourrait nuire à la lisibilité (visuelle ou imprimée) de l'adresse en question.

[89] Common Gateway Interface : Mécanismes qui permettent aux programmeurs du Web d'interagir avec des programmes s'exécutant sur un serveur HTTP.

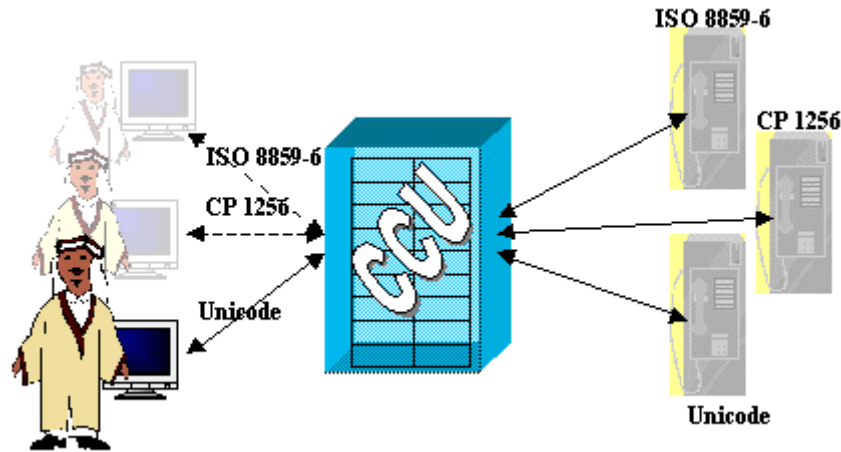


Figure 38 : Accès unilingue via codes de caractères unifiés

② Scénario 2 : travailler en mode client actif pour accéder à une ressource arabe via une URL latine ou arabe.

Étant donné que l'arabisation des applications informatiques sous-entend un bilinguisme arabe/latin (généralement arabe/français ou arabe/anglais), la ressource arabe à atteindre est identifiée soit par une URL arabisée soit par une URL latine au niveau du client selon un codage de caractères particulier (CCP = ISO 8859-1, ISO 8859-6, CP 1252, CP 1256). L'URL sollicitée par l'utilisateur est ainsi exprimée dans le CCP, traduite au niveau de la table de correspondance en URL mnémonique en CCU (i.e. Unicode) puis convertie en URL numérique. L'usage des deux possibilités arabe et latine dans l'identification d'une URL au niveau du client, sous entendrait que la ressource en question subit l'une des deux opérations suivantes :

- ❑ soit être déclarée au niveau de la table de correspondance dans deux versions linguistiques différentes pour correspondre au choix de l'utilisateur. Cette supposition est lourde en terme d'encombrement des tables de correspondance par des doublons de références aux mêmes ressources ;
- ❑ soit par traduction de l'une des deux versions linguistiques vers l'autre par un programme supplémentaire de traduction externe qui transmettrait en conséquence une seule version linguistique à la table de correspondance qui la convertirait en CCU. Cette hypothèse est également lourde voire même impossible étant donné que la traduction d'adresses URLs n'obéit pas à des règles sémantiques ou syntaxiques strictes. Une adresse URL en langue arabe ne peut pas être traduite en langues latines (et vice versa) sur des bases d'équivalence lexicales strictes. Par contre la solution de la translittération a été pour longtemps appréhendée comme solution possible pour convertir de l'ASCII aux écritures locales. La solution a fonctionné pour ceux qui maîtrisent bien l'écriture locale. La problématique de cette solution était la non disponibilité de programmes adéquats de translittération (illimités en nombres de paires linguistiques) à tout moment de consultation des documents virtuels. La restitution s'effectuait alors sous forme de caractères graphiques illisibles. Une démarche possible, mais très improbable, pour la concrétisation de cette alternative nécessiterait une conformité absolue au niveau des serveurs et des clients en langue arabe pour l'usage d'une version linguistique unique des ressources arabes. Cette version linguistique sera celle transmise à la table de correspondance au moment de la formulation de la requête.

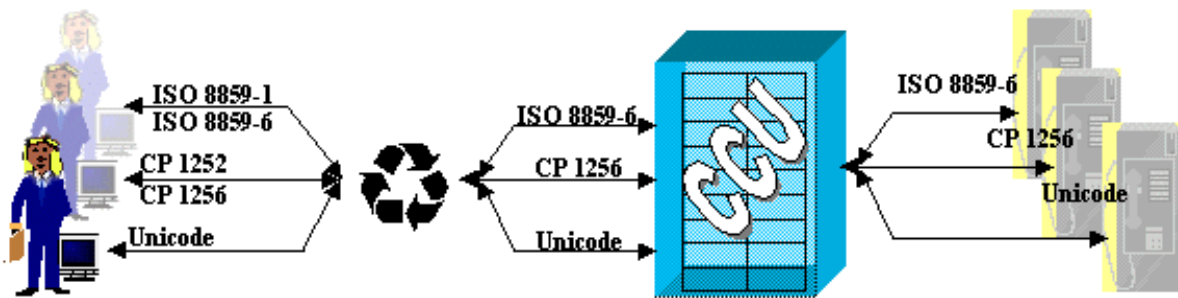


Figure 39 : Accès à une ressource arabe via une URL latine/arabe

③ **Scénario 3 : travailler en mode client actif pour accéder à une ressource latine via une URL arabe ou latine.**

Il s'agit du modèle précédent avec inversion des mécanismes de traduction et de conversion entre les versions linguistiques.

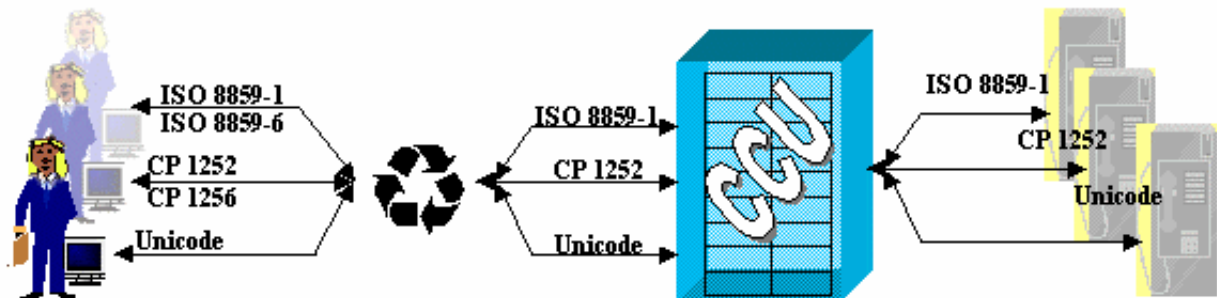


Figure 40 : Accès à une ressource latine via une URL latine/arabe

④ **Scénario 4 : travailler en mode client passif pour accéder à une ressource latine via une URL arabe ou latine.**

En mode client passif, l'URL d'une ressource d'information est par contre dynamique (URL de recherche, résultant d'un traitement de formulaire de requête ...). La passivité de l'utilisateur dans ce contexte est de ne pas contribuer à la saisie des URLs qu'il n'a pas besoin d'ailleurs de maîtriser car elles sont plutôt créées, codées et orientées vers les systèmes d'indexation et de recherche (moteurs, répertoires, robots...) qui en tiennent compte pour déclencher un processus de traitement interne.

En effet, la soumission d'un formulaire de requête génère une URL dont une partie, connue sous le nom de chaîne (ou partie) de requête (introduite par un point d'interrogation), reflète la chaîne de caractères saisie arbitrairement dans les zones utilisateur du formulaire et traduite en codes compréhensibles par le navigateur et les serveurs concernés.

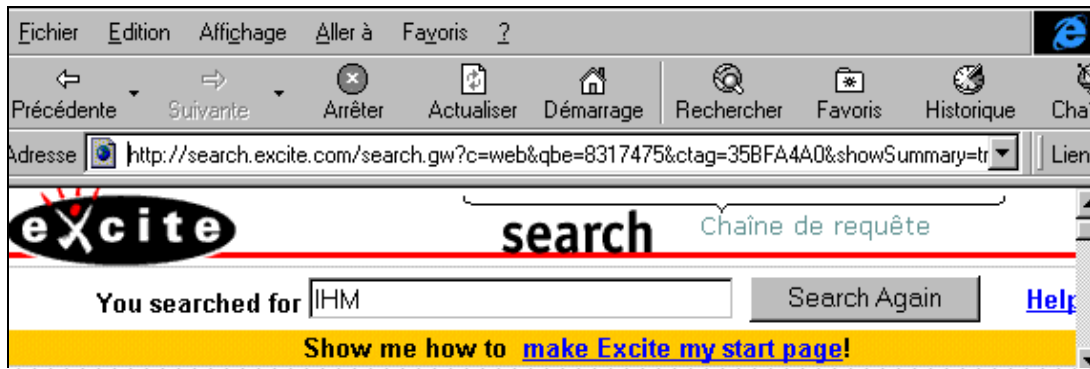


Figure 41 : La partie requête dans une URL dynamique

Les serveurs ont généralement de problèmes avec cette partie car elle contient des codes de caractères en dehors du répertoire ASCII non identifiés par les normes en cours d'usage. Les utilisateurs sont toujours enclins à exprimer leurs stratégies de recherche dans un langage naturel dans leurs langues d'origine. Les zones de saisie et les fenêtres d'options dans les formulaires doivent normalement accepter les données dans la langue de l'utilisateur avec la possibilité, au moment de l'envoi de la requête, de convertir tout le texte saisi dans un format que l'analyseur de la requête saura comprendre.

Le handicap majeur pour la requête par formulaire est le fait que les serveurs sont majoritairement basés sur un codage ASCII au niveau de leurs bases de données, systèmes caches et systèmes de fichiers. Si la restriction actuelle imposée pour l'usage exclusif de l'ASCII est levée, il serait indispensable de mettre en place un mécanisme d'identification des caractères utilisés lors de l'envoi de la requête au serveur afin que celui-ci (ou le client) puisse les convertir en ASCII standard. Nous n'avons décelé jusqu'à présent aucune référence d'un mécanisme capable de contrôler cette conversion en dehors du mécanisme de conversion des caractères Latin-1 vers l'ASCII. Les investigations ont démontré que très peu de serveurs gèrent des requêtes dans des langues basées sur des codes de caractères autres que l'ASCII^[90]. Des solutions de multilinguisme arabe/latin sur Internet sont pourtant déjà sur la route de la standardisation. Les produits de la société Alis Technologies^[91] (i.e. le kit Tango), la société Sakhr^[92] (i.e. le moteur de recherche Al Idrisi^[93]), la société AccentSoft^[94] (i.e. le navigateur Navigate With Accent) fournissent déjà un exemple de services de recherche en arabe intégral.

L'aspect multilingue du mode serveur

L'autre façon d'aborder la fonctionnalité de notre proposition pour un système multilingue est de voir son adéquation avec l'architecture client-serveur.

Dans les exemples précédents (modes clients actifs et passifs), nous avons analysé des situations du point de vue des clients générateurs d'URLs, donc influents directs ou indirects sur la nature des solutions offertes. La partie serveur, par contre est censée traiter des requêtes et des réponses en intrant soumises et transmises par des analyseurs et des serveurs de noms externes.

[90] Voir F. YERGEAU. A world-wide World Wide Web. Alis Technologies.

[91] Alis Tec. : <http://www.alis.com>

[92] Sakhr Ltd. : <http://www.sakhr.com>

[93] IDRISI = Intelligent Document Retrieval and Information Search for the Internet. <http://www.alidrisi.com/>

[94] AccentSoft. : <http://www.acenntsoft.com>

En tant que serveur, notre modèle est censé fonctionner à la manière des systèmes actuels spécifiés par les RFC 1034 et 1035 sauf que la table de correspondance devient la pièce maîtresse du système en question.

Une requête est soumise en mode numérique par un analyseur externe qui la soumet à la table de correspondance où elle est convertie en URL mnémonique par conversion des trois zones de l'URL comme précisé auparavant. La zone protocole est traduite en un identifiant mnémonique par la routine incorporée à la table de correspondance ; le nom de domaine est validé par le serveur de nom local et l'identifiant de la ressource est traité en rapport avec le nom du domaine auquel il est rattaché. Deux situations sont alors possibles :

- ❑ Si le domaine et la ressource sont repérés sur le serveur local, la table de correspondance effectuera une conversion de leur forme numérique vers une forme mnémonique en Codes de Caractères Unifiés (CCU). Cette version est ensuite retraduite en Codes de Caractères Particuliers (CCP) qui permettra de traiter la requête en local et de rapatrier les ressources stockées sur les serveurs locaux de données. La transmission de la réponse prendra le chemin inverse en passant par toutes les opérations de conversion et de traduction précédentes.
- ❑ Si le nom du domaine et l'identifiant de la ressource sont des éléments externes, la table de correspondance les expédiera en requête numérique vers l'analyseur principal qui les traitera en s'adressant à des serveurs de noms externes. Le processus prendra le temps nécessaire et les chemins virtuels adéquats jusqu'à résolution complète de la requête qui sera ensuite retransmise à la table de correspondance en version numérique. Celle-ci en avisera l'analyseur externe qui a émis en début de parcours la requête en question.

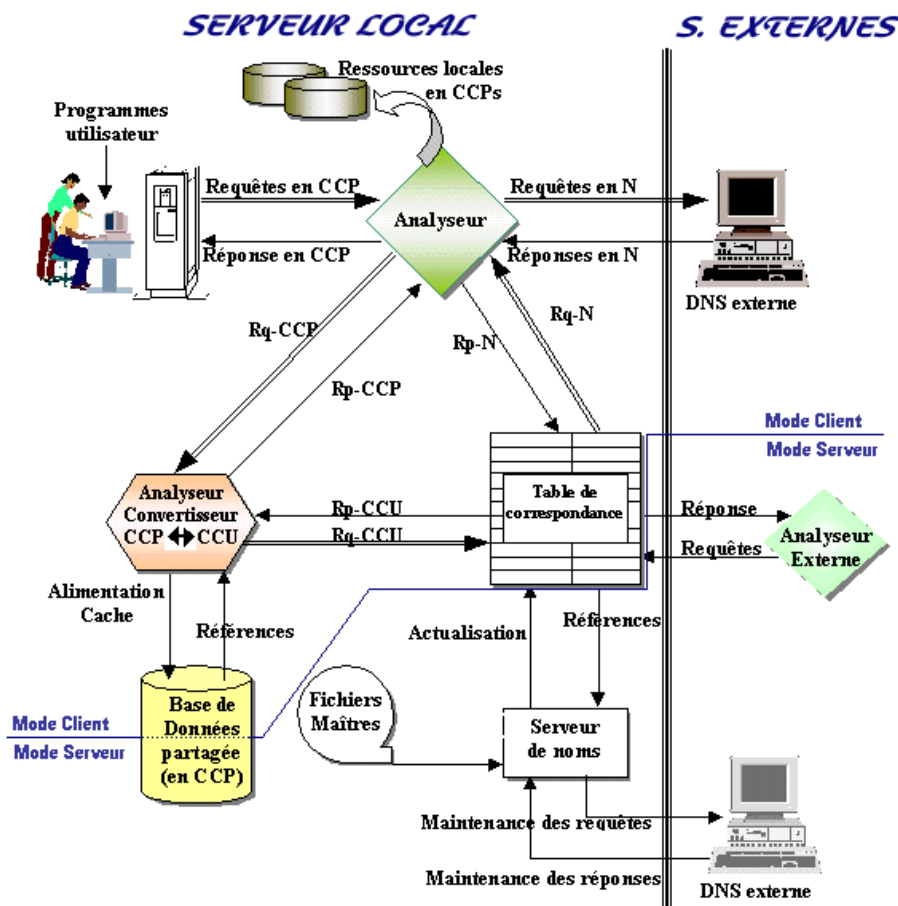


Figure 42 : Fonctionnement d'un modèle DNS multilingue

En parallèle, le serveur de noms local continuera en permanence les deux opérations de maintenance des requêtes et des réponses auprès des serveurs de noms externes. La base de données partagées continue de son côté à jouer le rôle de système cache local et de cumulatif en CCP des références d'actualisation accomplies par le serveur de noms local.

Parmi les points qui restent à définir de manière plus profonde dans cette proposition, le choix des jeux des caractères codés à utiliser soit au niveau des CCP soit celui des CCU. Ce dernier est le plus important car il jouera le pivot central de toutes les opérations de conversion dans tous les systèmes multilingues. Ce drainage vers une solution uniforme et univoque devrait répondre à cette réclamation d'une universalisation équitable et impartiale. Déjà, le rôle d'Unicode et de ISO/IEC 10646 se profile comme la solution incontournable. La littérature courante signale un début de consensus vers l'internationalisation des URLs et des DNS sur la base d'UTF-8 (RFC 2040) qui constitue la forme de transformation du jeu de caractères universel sur 8 bits (§ [III-3.4.1.3.2](#)).

4.2.1.2.3 La transformation UTF-8 et le format de code %HH

L'une des solutions de codage des caractères à 8 bits a été proposée par Martin Dürst^[95] et François Yergeau^[96] afin d'étendre cette limite linguistique des noms des domaines et des URLs aux caractères non-ASCII voire même les caractères non latins codés sur 8 bits. Leurs propositions concernent la transformation des codes des caractères URL vers le format de transformation 8-Bits : UTF-8. Le jeu des caractères UTF-8 couvre la majorité des langues écrites du monde permettant ainsi d'appliquer toutes les alternatives linguistiques de la localisation. « *UTF-8 is the recommended character set for protocols that are evolving beyond ASCII [RFC2130]* »^[97].

a) Le format %HH

Il est à rappeler à ce stade que les différentes structures de l'adressage sur Internet incluent des caractères codés en séries d'octets traduites en URLs selon un mécanisme de codage particulier. Le codage des URLs, comme spécifié dans la RFC 1738, est le processus par lequel un seul octet dans la représentation d'une URL pourrait être étendu, quand nécessaire, vers trois octets à valeurs ASCII dans un format « %HH » où « HH » est la valeur de l'octet en hexadécimal. Ce mécanisme est direct et réversible uniquement pour un sous-ensemble de caractères ASCII, en l'occurrence les caractères URLs. Toutes les autres valeurs d'octets, surtout celles qui représentent des caractères réservés ou non sûrs (l'espace, la barre verticale ...) ainsi que les caractères en dehors du répertoire ASCII, sont obligatoirement codées en format %HH. L'avantage de cette solution est que l'équivalent ASCII des caractères non-ASCII permettrait à ceux-ci de passer les barrières des nœuds exclusivement 7 bits sans subir une élimination de leur 8^e bit, d'être affichés sur écran et d'être saisis au clavier en formats codés sur des sites ou leurs formes graphiques ne sont pas disponibles.

Pour illustrer le fonctionnement de cette méthode, l'équivalent UTF-8 du terme « Français », transcrit en Unicode, serait « Fran%c3%e7ais ». Puisque le caractère « ç » n'appartient pas au répertoire US-ASCII, il est converti vers son équivalent hexadécimal en UTF-8 sous forme de

[95] Martin DÜERST. Internationalization of Domain Names. University of Zurich. Décembre 1996. <http://www.ifi.unizh.ch/groups/mml/mduerst/urli18n.html>

[96] François YERGEAU. Internationalization of URLs. Alis Technologies. Septembre 1996. <ftp://ds.internic.net/rfc/rfc2044.txt> ou <http://www.alis.com:8085/~yergeau/url-00.html>

[97] Stuart KWAN, James GILROY. Using the UTF-8 Character Set in the Domain Name System. Internet Draft. Microsoft Corp. Mars 1998. <http://draft-skwan-utf8-dns-01.txt>

deux Octets « c3 » et « a7 ». La déclaration de cette conversion est effectuée au sein de l'URL grâce au symbole « % ».

Dans l'exemple suivant, les caractères accentués transcrits en ISO/IEC 8859-1 dans une adresse URL sont convertis en codes « %HH » composés de caractères ASCII, donc valables pour transiter par les nœuds et les serveurs à 7 bits sur Internet. L'utilitaire de restitution se chargera de reconvertir ces codes à leurs formes originales.

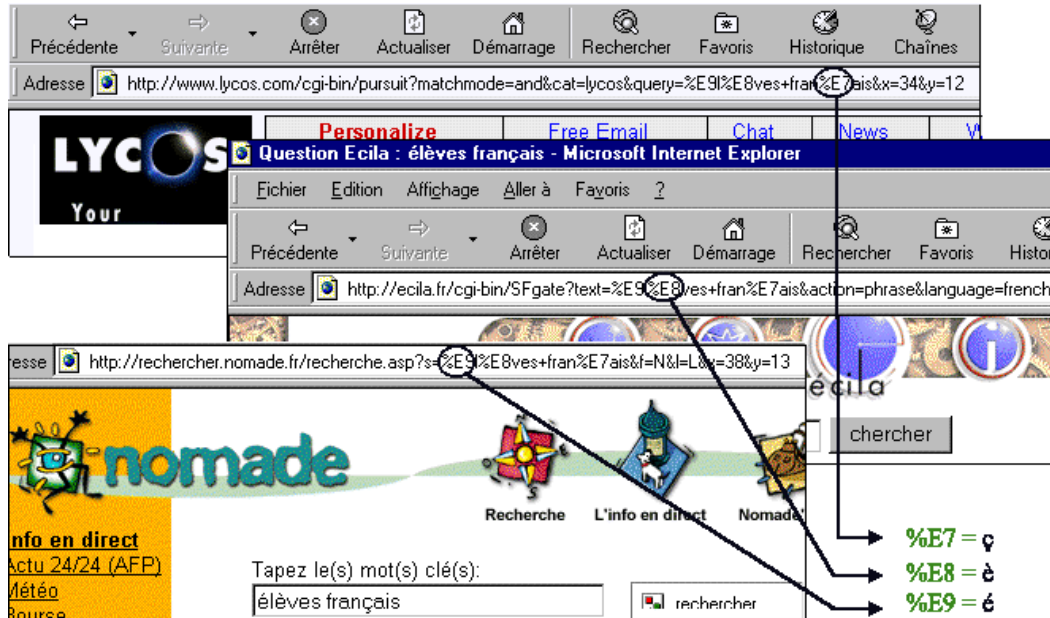


Figure 43 : La conversion des caractères non-ASCII dans les URLs

Dans cet autre exemple, une équation de recherche est exprimée en caractères arabes et latin étendu par le moteur de recherche Metacrawler puis convertie, dans la zone requête de l'URL du navigateur Tango multilingue d'Alis Technologies, en équivalents Hexadécimaux là où c'est nécessaire (Non ASCII).

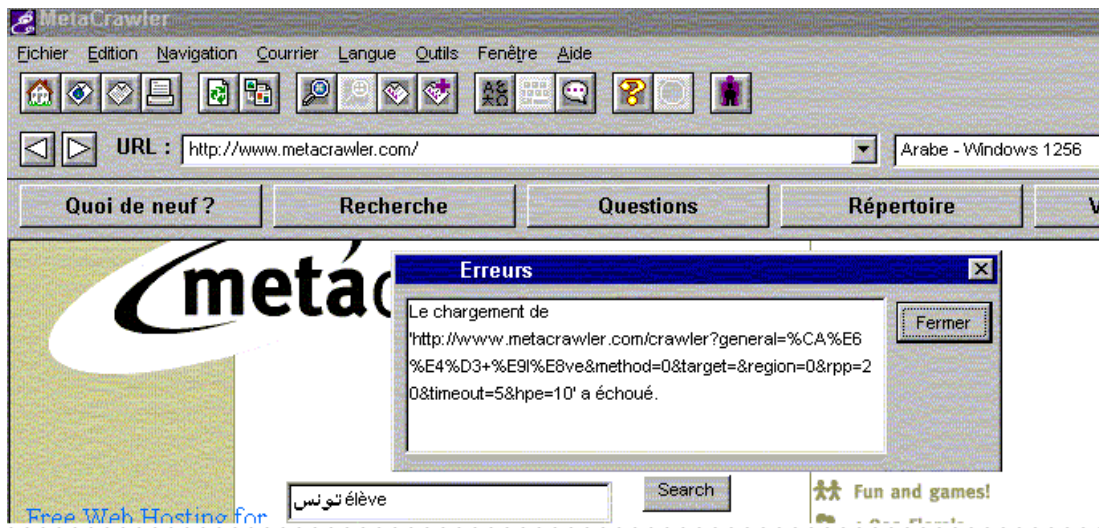


Figure 44 : Le codage hexadécimal des caractères non ASCII au niveau des URLs

Cette pratique est devenue courante auprès des agents utilisateurs traitant des URLs non-ASCII. Son inconvénient majeur est l'absence d'une indication du jeu de caractères utilisé pour désigner

les caractères non-ASCII codés en « %HH ». Selon J.Bourbonnais et F.Yergeau, le code « %E9 » peut à la fois correspondre au caractère latin « é » ou au caractère cyrillique « IOU » ou à beaucoup d'autres caractères dans d'autres écritures^[98]. Selon Gavin T. Nicol, de l'*Electronic Book Technologies-Japan*, quand un utilisateur japonais reçoit un formulaire américain (conçu sur une base ASCII), le remplit en caractères japonais puis le renvoie au serveur émetteur, l'URL transmise au serveur émetteur peut avoir la forme suivante :

<http://foo/bar?baz=%B5%6B%85%9A>

Nous avons expérimenté la même chose avec une requête par chaîne de caractères arabes posée au moteur Metacrawler via le navigateur latin MSIE Explorer 4.0 tournant sous Arabic Windows95.

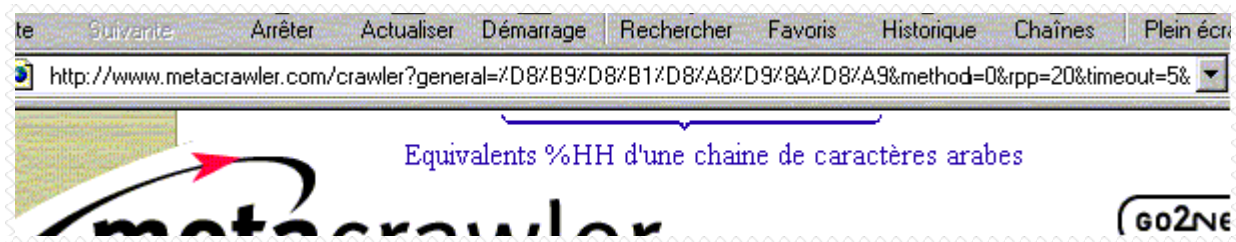


Figure 45 : Conversion intégrale en %HH d'une chaîne de caractères arabes dans une URL de requête

Le problème de cette forme d'adresse surgirait au moment du décodage de l'URL par le serveur traitant la requête : puisque celui-ci n'a aucune indication sur le jeu de caractères codés ou la technique de codage utilisée, le système ne pourra pas facilement reconnaître le texte après le point d'interrogation (qui annonce la partie requête d'une URL) ou l'adresse entière^[99].

Du côté des serveurs, la problématique concerne à la fois la réception et l'envoi de chaînes de caractères. Le grand problème des caractères reçus par un serveur est le fait qu'une grande majorité de ces caractères proviennent par le moyen des formulaires de requête où l'indication des jeux de caractères utilisés n'est pas normalisée. En ce qui concerne les données émises, l'un des problèmes majeurs auxquels font face les serveurs est l'attitude à prendre face à une requête d'un utilisateur exprimée dans un encodage de caractères incompréhensible. Dans ce cas de figure, HTTP prévoit la solution suivante : « *Quand une ressource est disponible uniquement dans un jeu de caractères autre que celui par défaut, et que le jeu de caractères n'est pas listé dans le champ « Accept-Charset », il n'est possible au serveur d'envoyer l'entité (!) que si le jeu de caractères pourrait être identifié à travers un paramètre « Charset » approprié au type de média ou à l'intérieur du format du type de média lui-même* » (Trad.)^[100].

Le codage en format %HH s'effectue généralement en 2 étapes :

1. Première étape : conversion des caractères potentiellement utilisés pour la définition des URLs dans une séquence de caractères codés en UTF-8 ;
2. Deuxième étape : conversion de tous les caractères non ASCII de l'URL vers un format « %HH » où « HH » sont les valeurs hexadécimales du caractère non ASCII.

[98] Jean BOURBONNAIS, François YERGEAU. Languages on the Internet. Alis Technologies inc. Montréal. http://www.crim.ca/inet96/papers/a5/a5_3.htm

[99] Gavin T. NICOL. The Multilingual World Wide Web. Electronic Book Technologies. Inc. <http://www.sil.org/sqml/nicol-multwww.html>

[100] Traduction du texte de Gavin T. NICOL. Gavin T. Nicol. The Multilingual World Wide Web. Electronic Book Technologies. Inc. Op.Cit.

Selon Martin Dürst^[101], deux possibilités se présentent si l'on se propose d'utiliser les URLs à caractères 8 bits :

1. Soit interpréter tout l'URL directement comme séries d'Octets puis le convertir au format %HH. Cela suppose que l'URL n'est pas transcodé à l'origine ;
2. Soit interpréter l'URL comme caractères, le convertir en UTF-8 puis en format %HH. Cela suppose que le serveur utilise UTF-8 pour encoder les caractères.

Une troisième solution est envisageable mais non recommandée consiste à combiner les deux hypothèses précédentes : une URL transcodée et transmise à un serveur qui ne supporte pas UTF-8 serait une situation délicate à résoudre.

b) La solution Unicode

Le questionnement que nous pouvons toutefois adresser à ce niveau concerne l'étendu de l'I18n de l'UTF-8 à tout le répertoire UCS2 et UCS4. En d'autres termes, en l'absence d'une indication du jeu de caractères codés au niveau de l'URL, la majorité des systèmes utilisent le jeu de caractères codés implicite du navigateur en cours d'utilisation. La situation actuelle appliquée à la majorité absolue des transactions le codage ISO/IEC 8859-1. Les caractères codés en format « %HH » sont donc implicitement codés et décodés selon leurs équivalents Latin-1 ([Fig.45](#)).

Or dans un contexte multilingue, on aurait besoin d'un format de codification qui permettrait de garder la valeur sémantique des caractères non-ASCII transportés. Il n'existe aucune norme capable de représenter correctement des caractères arabes ou non-latins en anglais ou en ASCII. Même les solutions de translittération, (i.e. la norme ISO 233^[102]) ne répondent pas aux besoins de l'I18n des URLs car elles sont énormément basées sur les diacritiques et les digraphes complexes. La question serait alors de savoir comment satisfaire à travers l'usage d'UTF-8 ceux qui auraient besoin d'utiliser des URLs dans leurs propres langues ?

La solution serait de faire usage d'un seul jeu de caractères vers le quel le jeu local serait converti au moment du transfert. Les caractères UCS (ISO/IEC 10646) dans leur format de transformation UTF-8 serait dans ce sens une bonne solution pour coder les caractères à 8 bits en format %HH là où 8 bits ne seraient pas admis.

Unicode, comme présenté auparavant dans ce chapitre ([§ III-3.4.1](#)), offre un avantage particulier sur toutes les autres alternatives de codage des caractères : c'est un jeu de caractères unique avec des possibilités de conversion bidirectionnelles (round trip) de et vers les autres formes de codages de caractères en 7, 8 et 16 bits. Avec Unicode en format UTF-8, le codage et le décodage des caractères en format %HH seraient conformes à leurs valeurs d'octets dans la chaîne Unicode.

Larry Masinter et Martin Dürst expliquent dans un avant projet Internet (Internet-Draft) que, pour définir des URLs multilingues, les séquences des caractères qui incluent des caractères non-

[101] Martin DÜERST. Archives du forum « URL Internationalization ». Lundi 24 février 1997. <http://www.acl.lanl.gov/URI/archive/uri-96.messages/0010.html>

[102] Information et documentation – Translittération des caractères arabes en caractères latins. Partie 2 : Langue arabe – Translittération simplifiée. ISO 233-2 : 1993 (F) 6p.

ASCII devraient être transcrites, avant leur inclusion comme éléments URI, selon la méthode suivante^[103] :

- ❑ Représenter les caractères comme une séquence de caractères ISO/IEC 10646 ;
- ❑ Normaliser les séquences de caractères pour réduire l'ambiguïté ;
- ❑ Coder le résultat en UTF-8 (RFC 2279^[104]) ;
- ❑ Utiliser le format hexadécimal %HH (RFC 2396) pour coder tout octet qui ne correspond pas à un caractère autorisé ou non réservé.

Cette syntaxe correspond à la définition de la syntaxe des URIs génériques (RFC 2396), des URN (RFC 2141) et des définitions récentes du schéma URL (RFC 2192 et RFC 2384).

Quand un système quelconque reçoit une URL en arabe ou en Devanagari ou une requête par formulaire d'un moteur de recherche dans le format %HH, aucun problème d'écriture, ni de saisie ni d'impression ne serait alors observé. Le plus important est que si les URLs sont transmises en mode UTF-8 (avec l'encodage %HH quand nécessaire) le système local peut identifier le caractère référencé par les octets codés dans l'URL. En conséquence, il affichera les caractères qu'il est capable de reconnaître directement et affichera ceux qu'il ne parvient pas à identifier en format %HH. Par cette méthode, les URLs garderaient leur aspect convivial malgré la présence de caractères non-ASCII. Le problème est maintenant situé au niveau de l'unanimité d'usage d'Unicode par les concepteurs de logiciels et les constructeurs de matériels pour favoriser la mise en place de cette solution.

Cette méthode commence déjà à faire son entrée auprès des fournisseurs d'accès Internet multilingue et plus particulièrement arabes. La société *Engineering Graphical Solutions – Ayna*^[105] à titre d'exemple a adapté son moteur de recherche *Nathra*^[106] à une grande variété de règles lexicales arabes pour la recherche avancée sur racine, par troncature et par lemmatisation. Son support des codages CP-1256, ISO 8859-6 et UTF-8 permet à tout utilisateur muni d'un support arabe minimum d'accéder à ses ressources arabes. Netscape Communicator et MSIE Explorer ont déjà migré vers le codage Unicode et nous retrouvons aujourd'hui beaucoup d'aisance au niveau du choix d'un format ou d'un autre à partir de l'interface utilisateur de ces navigateurs.

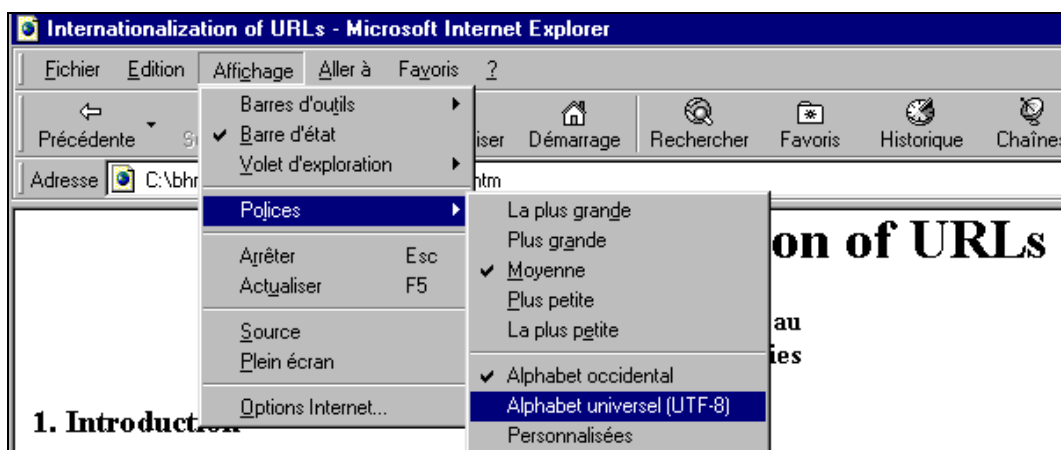


Figure 46 : Choix des codes des caractères sous MSIE Explorer 4

[103] Larry MASINTER. Martin DÜERST. Representing non-ASCII Characters in URIs and Extended URIs. Internet Draft. Août 1998.

[104] UTF-8, a Transformation Format of Unicode and ISO 10646. <http://www.imc.org/rfc2279>

[105] <http://www.ayna.com>

[106] <http://www.egsx.com/products/nathra/>

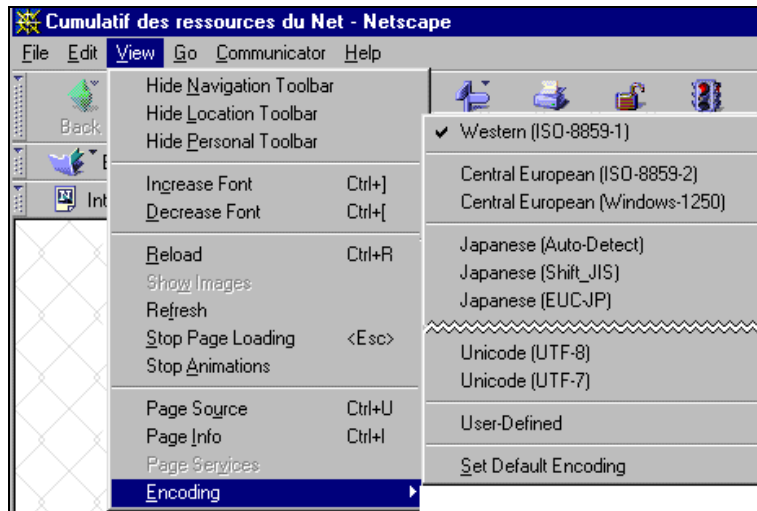


Figure 47 : Choix des codes des caractères sous Netscape Communicator

Par contre les autres solutions multilingues en arabe/latin du milieu commercial et de la recherche n'ont pas encore migré vers la solution Unicode et restent par conséquent limités aux normes de codage CP-1256 de Windows, ISO/IEC 8859-1 et ISIRI 3342 (pour le persan). À titre de référence, nous citons parmi cette catégorie ALM et AraMosaïc de *Langbox International*^[107] et PMosaïc de la *California-based Simorgh Software* (Voir également la [liste des produits](#) à la fin du chapitre premier).

c) La Bidirectionnalité

Un problème supplémentaire mais tout particulier aux langues à directionnalité Droite-Gauche (RTL) reste encore non résolu : la bidirectionnalité d'une URL multiscrits. En d'autres termes, comment une URL composée de plusieurs parties (protocole, DNS, requête) codées en plusieurs scripts à la fois pourrait-elle afficher les caractères sur une seule ligne de texte ?

Étant donné la complexité dans la gestion de ce genre d'URLs, les sites arabes actuellement en fonction avec des outils partiellement ou entièrement arabisés sur Internet, continuent à faire usage d'URLs ASCII bien qu'avec toutes les variantes Unicode et UTF-8. Trois problèmes sont adressés à ce niveau de complexité : la bidirectionnalité au niveau des différentes zones de l'adresse URL, l'étiquetage des jeux des caractères codés et la localisation de l'identifiant du protocole.

1. Les caractères arabes et hébreux sont écrits de droite à gauche. Leur mixture avec des caractères latins résulterait en une divergence entre les séquences logiques et graphiques de la même adresse (§ IV.2.3). La présence d'une zone protocole codée en ASCII au début de l'adresse URL (<http://>; <ftp://> etc.) et la présence d'une chaîne de caractères Droite/Gauche dans le reste du schéma convertirait automatiquement une URL en une chaîne de caractères bidirectionnelle.

Une URL arabe serait ainsi composée de caractères ASCII (protocole), de caractères arabes (ressource) et de caractères neutres (séparateurs) comme « / », « . », « : », « # » etc. La tendance de l'i18n est de donner la priorité absolue à l'agencement de l'URL conformément à la disposition des caractères de séparation avant le texte qui les sépare. Étant donné que les

[107] <http://www.langbox.com/alm/alm.html>

caractères ASCII sont souvent majoritaires et omniprésents grâce à l'identifiant du protocole, les « i18nistes » (!) recommandent que l'orientation globale de l'URL soit une orientation Gauche-Droite même dans des contextes à orientation totalement Droite-Gauche. Ils préfèrent que seules les zones individuelles d'un nom de domaine ou d'un URI ait une orientation Droite-Gauche. La zone DNS de l'URL «<http://isd.rnu.tn>» serait mieux présentée en langue arabe comme «<http://dsi.unr.nt>» (<http://تن.الجامعة.توثيق>) au lieu de «<http://nt.unr.dsi>» (<http://توثيق.الجامعة.تن>).



Figure 48 : URL multilingue arabe/latin (fictive !)

La présence des caractères neutres (« / », « . », « : », « # » etc.), compliquerait davantage la situation si l'affichage fait recours uniquement à l'algorithme BIDI d'Unicode conçu essentiellement pour du texte libre sans balises de formatage (i.e. HTML ou LaTeX). L'orientation d'une parenthèse survenue entre un caractère arabe et un autre latin n'est pas toujours évidente sans traitement explicite. Dans une partie précédente (§ III-3.4.2.2), nous avons précisé que les caractères neutres prennent les attributs des caractères qui les entourent et s'affichent dans une directionnalité relative. Deux valeurs supplémentaires sont alors introduites pour définir la propriété directionnelle d'un caractère neutre. La valeur « LRM » (Left to Right Mark) force la directionnalité Gauche-Droite et « RLM » (Right to Left Mark) attribue la propriété Droite-Gauche.

François Yergeau, cité par Martin Dürst^[108], propose que le rendu d'une URL multilingue devrait être précédé par une opération de forçage « LTR » aux caractères neutres (en les faisant entourer par une paire de marques « LRM ») afin que la bidirectionnalité soit appliquée uniquement au niveau du texte des différentes zones de la structure URL. Une URL en langue arabe (saisie en mode maître latin lui-même forcé par la position initiale du protocole codé en ASCII) serait donc saisie dans la logique de l'utilisateur par une évolution progressive Droite-Gauche à partir du nom du domaine jusqu'à la partie ressource. Le système accomplirait au fur et à mesure des inversions de position sur les parties de l'adresse séparées par une ponctuation forcée « LTR ». Cette pratique aurait à notre sens deux inconvénients majeurs :

- ❑ D'une part il n'y aurait plus de conformité entre la structure de l'URL ainsi conçue et son traitement dans un corps de texte (i.e. lien hypertexte) où le forçage « LTR » des caractères neutres ne s'applique pas ;

Zone URL <i>Marque "LRM"</i>	http://وثيقة/ندوة/تن.الجامعة.المعهد 1 2 3 4 5
Zone Texte <i>Marque "RTL"</i>	http://المعهد.الجامعة.تن/ندوة/وثيقة 5 4 3 2 1

Figure 49 : Nuances entre marques « LRM » et « RTL » sur zones URL et texte

[108] Martin DÜERST. Archives du forum « URL Internationalization ». Lundi 24 février 1997. <http://www.acl.lanl.gov/URI/archive/uri-96.messages/0010.html>

- D'autre part, en mode interne logique du backstore (§ IV 2.3.1), les caractères en orientation Droite-Gauche auraient une disposition inverse à celle de leurs équivalents en mode texte « LTR » (Tab.28). Ceci est d'ordre à empêcher les opération de copier/coller entre URL et zone texte et d'empêcher la recherche du même mot dans les deux zones à la fois. Cette difficulté serait dépassée si l'utilisateur disposerait d'une possibilité de traiter les URL dans les deux modes « RTL » et « LTR ». Nous récapitulons ces conflits de directionnalités entre modes visuel et logique entre directions « LTR » et « RTL » d'une part et entre marques « LRM » et « RLM » de l'autre à travers le tableau suivant :

Contexte	URL Latine	URL Arabe
Mode visuel LTR	http://isd.rnu.tn	http://المعهد.الجامعة.تن
Mode Visuel LTR + LRM	http://isd.rnu.tn	تن.الجامعة.المعهد
Mode logique LTR	http://isd.rnu.tn	ن ت . ة ع م ا ج ل ا . د ه ع م ل
Mode logique LTR + LRM	http://isd.rnu.tn	ن ت . ة ع م ا ج ل ا . د ه ع م ل
Mode visuel RTL	http://isd.rnu.tn	http://المعهد.الجامعة.تن
Mode visuel RTL + RLM	tn.rnu.isd//:http	http://المعهد.الجامعة.تن
Mode logique RTL	nt.unr.dsi//:ptth	ن ت . ة ع م ا ج ل ا . د ه ع م ل
Mode logique RTL + RLM	nt.unr.dsi//:ptth	ن ت . ة ع م ا ج ل ا . د ه ع م ل

Tableau 28 : Nuances entre marques « LRM » et « RTL » dans le traitement logique des URL bilingues

La proposition de F.Yergeau, en entourant l'URL dans une zone de texte « RTL » par une paire de marques « LRE-PDF » (§ III-3.4.2.2), fait face, elle aussi, à des incohérences attribuées au changement de positions des zones de l'URL suite au traitement des programmes concernées de l'URL comme bloc de texte normal. Faire en sorte que l'URL aurait une directionnalité intégrale et forcée de gauche à droite, puis entourée par une paire de marques « LRO-PDF », aboutirait au même conflit entre la représentation logique (en backstore) des deux modes URL et texte.

- La multiplicité des codes de caractères utilisés pour les différentes parties de l'URL suppose que le programme qui gère l'interface utilisateur est censée reconnaître et identifier toutes les solutions d'i18n des différentes parties d'un schéma URL. Ce qui est loin d'être réalisable pour le moment étant donné que certaines des solutions de l'i18n actuellement utilisées ne sont pas documentées ni même connues. D'autre part, un étiquetage linguistique de chaque partie de l'URL encombrerait l'adresse, alourdirait son transfert et la rendrait moins lisible au moment de son impression.
- Notre critique à cette résignation devant la légitimité incontestée de l'identifiant exclusivement ASCII du protocole émane de plusieurs facteurs :
 - Les navigateurs de dernières générations proposent d'omettre la partie protocole de l'URL et de l'introduire automatiquement après saisie du reste de l'adresse et validation par l'utilisateur (i.e. Option « Utiliser la saisie semi-automatique » sous MSIE Explorer 4);

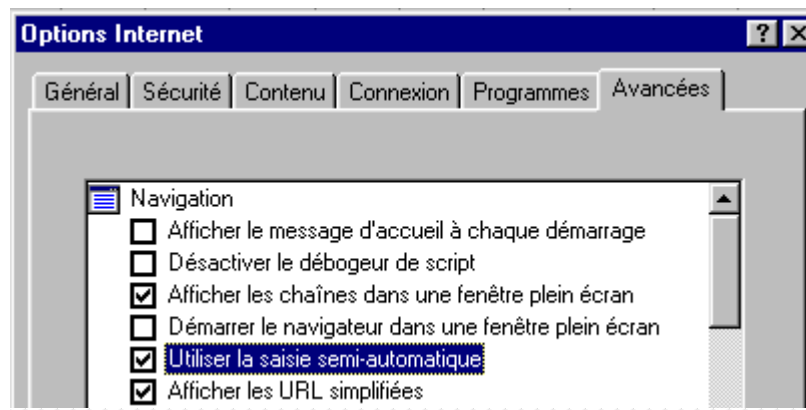


Figure 50 : Paramétrage d'URLs sans identifiants des protocoles sous MSIE4.

- ❑ Comme nous l'avons souligné dans notre proposition d'i18n d'URIs, la partie protocole pourrait être soumise à un programme de traduction au niveau du navigateur local pour l'adapter à la version linguistique de l'utilisateur. Elle pourrait aussi faire objet d'un menu déroulant dans lequel l'utilisateur choisirait dans une liste de choix le protocole qu'il voudrait utiliser. Derrière la couche interface, le système réadaptera le code protocole au code système en usage. Certains produits, parmi lesquels nous avons déjà donné l'exemple de l'éditeur HtmlEdit (Fig.51), font usage de modèle d'interface pour saisir le protocole à définir pour un lien Hypertexte. MS Word 97 en est un autre.

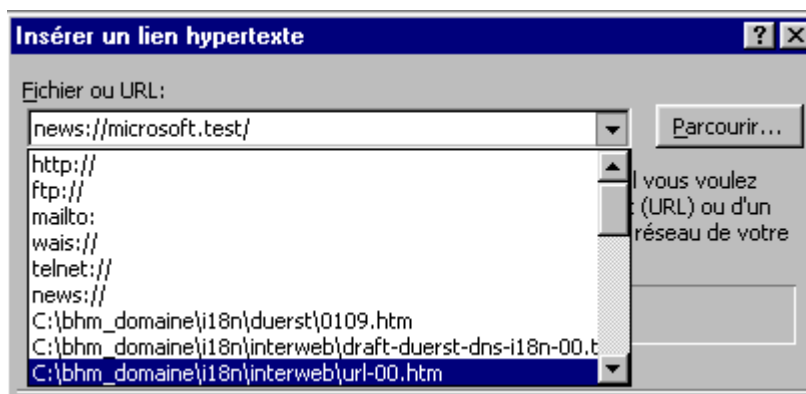


Figure 51 : Liste de choix de protocoles sous MS Word 97

Le dernier obstacle vers l'i18n et la l10n^[109] intégrales des URLs pourrait ainsi donc être surmonté. Seulement, des prédispositions essentielles devraient être remplies:

- ❑ une convergence vers un jeu de caractères codés universel unique et univoque vers lequel tous les jeux des caractères codés locaux seraient traduits en amont et en aval du processus de communication ;
- ❑ une uniformisation des interfaces Homme-Machine dans le traitement des schémas d'adresses URLs (étiquetage et négociation linguistiques) ;

d) Les pronostics

Une série de questions restent encore valable pour la situation actuelle de l'état de l'art en matière de traitement des URLs multilingues :

[109] Localisation = l + 10 caractères + n = L10n.

1. Faudrait-il que les URLs incluent des méthodes normalisées pour le codage multiscrits ?
2. Si oui, devrait-il se faire sur la base du format %HH du codage UTF-8 d'Unicode ?
3. Faudrait-il tenir compte des jeux des caractères nationaux ainsi que de l'ISO 2022 comme alternatives locales à l'Unicode ?
4. Si oui, est ce qu'il y aurait une obligation de baliser les URIs par des identifiants de jeux de caractères codés ?
5. Est-il évident que les nouveaux standards allaient assurer la compatibilité rétrospective et que les navigateurs seraient conçus, comme ils le sont aujourd'hui, pour supporter plusieurs types de comportements utilisateurs non normalisés ?
6. Est-ce qu'un prochain standard pourrait fournir des URLs Unicode à la source ?

La réponse à ces questions émergera progressivement des laboratoires de recherche, des expériences des universitaires et des académiciens ainsi que du contexte commercial. Mais à notre sens ceci prendra un peu plus de temps car le traitement des URLs multilingue nécessite un soubassement technologique plus complexe que celui de l'i18n des protocoles de transfert (HTTP) et de balisage hypertexte (HTML) des documents multilingues. L'étude de ces deux protocoles passe obligatoirement par une revue critique des deux aspects les plus déterminants du multilinguisme des documents HTML : d'une part l'étiquetage et la négociation de la langue des documents entre clients et serveurs conformément au protocole HTTP (Hypertext Transfer Protocol) et d'autre part la gestion de la bidirectionnalité dans le rendu des textes codés en HTML. Ce sont les deux éléments qui restent à étudier dans cette partie consacrée à la portée multilingue des protocoles d'Internet.

4.2.2 Les spécificités HTTP au profit du multilinguisme d'Internet

Sans revenir sur les particularités du balisage HTML, (Voir pour cela le texte du RFC 2070 cité en référence plus haut ou aller au § [III-4.2.3](#)) ni celles du protocole de transport HTTP (voir pour cela le texte du RFC 2068 cité en référence plus haut), les deux mécanismes de transfert et de balisage sont définis dans la structure interne d'un document HTML au niveau des deux parties principales : l'en-tête (HEAD) et le corps de document (BODY).

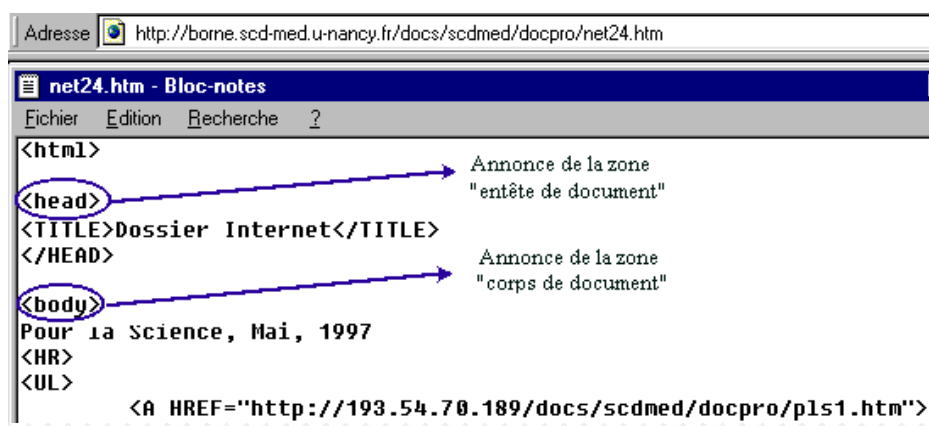


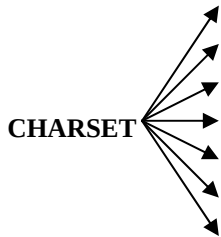
Figure 52 : Structure simplifiée d'un document HTML

Nous analyserons dans cette partie les aspects les plus révélateurs de la dimension multilingue du protocole HTTP. Nous en présenterons également une approche critique des différentes propositions en matière d'i18n d'Internet. Nous terminerons par une proposition personnelle relative à la problématique en question avec un regard particulier sur la langue arabe et sa cohabitation avec les langues latines. Nous excluons dans cette partie les particularités inhérentes à la gestion de la bidirectionnalité qui fera l'objet de la partie suivante prévue pour le balisage BIDI de HTML.

Le protocole HTTP est entré en application sur Internet en 1990 dans sa version 0.9. Il s'agissait d'un simple protocole de transfert de texte brut à travers Internet. La version HTTP 1.0 (RFC 1945)^[110] a été améliorée pour supporter un codage conforme aux modèles d'en-têtes MIME (Multipurpose Internet Mail Extensions).

4.2.2.1 La solution MIME

HTTP 1.1 (RFC 2068)^[111] gère de façon plus efficace les deux procédures de l'encodage des caractères multilingues et la négociation de la langue des documents au moment de leur transfert. Les deux opérations sont respectivement nécessaires pour l'interopérabilité et l'i18n des sites Web. Face à la diversité linguistique des ressources Internet, HTTP attribue au paramètre « Charset » de l'en-tête HTTP/MIME les valeurs des jeux des caractères codés suivantes :



"US-ASCII"		
"ISO-8859-1"	"ISO-8859-2"	"ISO-8859-3"
"ISO-8859-4"	"ISO-8859-5"	"ISO-8859-6"
"ISO-8859-7"	"ISO-8859-8"	"ISO-8859-9"
"ISO-2022-JP"	"ISO-2022-JP-2"	"ISO-2022-KR"
"UNICODE-1-1"	"UNICODE-1-1-UTF-7"	"UNICODE-1-1-UTF-8"
Token		

Tableau 29 : Les valeurs du paramètre « CHARSET »

Le paramètre "charset" identifie le jeu de caractères utilisé dans la définition d'un certain nombre de types de documents (généralement textuels). En l'absence de ce paramètre dans des documents reçu par HTTP, la valeur par défaut du jeu de caractère "ISO-8859-1" leur est automatiquement attribuée. Bien que HTTP ait évolué vers la valeur Unicode par défaut, son usage n'est pas encore entièrement généralisé. Il est par conséquent recommandé que toutes les données codées dans un jeu de caractères autre que ISO 8859-1 soient explicitement référencées par la valeur exacte de son propre jeu de caractères.

HTTP est également utilisé comme protocole générique pour communiquer entre les agents utilisateurs sur différents systèmes supportés par SMTP (Simple Mail Transfer Protocol), NNTP (Network News Transfer Protocol), FTP (File Transfer Protocol), Gopher et WAIS.

[110] T. BERNERS-LEE, R. FIELDING, and H. FRYSTYK, « Hypertext Transfer Protocol -- HTTP/1.0. », RFC 1945 MIT/LCS, UC Irvine, May.1996 <http://www.ics.uci.edu/pub/ietf/http/rfc1945.html>

[111] R. FIELDING. UC IRVINE. J. GETTYS. J. MOGUL. H. FRYSTYK. T. BERNERS-LEE. Hypertext Transfer Protocol -- HTTP/1.1. Request for Comments: 2068. Network Working Group. Janvier 1997. <http://www.ics.uci.edu/pub/ietf/http/rfc2068.txt> ou <http://ds.internic.net/rfc/rfc2068.txt>.

La particularité qui caractérise HTTP par opposition aux autres protocoles Internet de traitement et de stockage des données est sa stricte relation avec le mécanisme de transmission et d'échange des données entre systèmes ouverts et distribués. En effet, si les problèmes d'interprétation des séquences d'octets ont pu être résolus par plusieurs façons sur des systèmes différents, la transmission des séries d'octets entre plates-formes et systèmes hétérogènes nécessiterait par contre la définition d'une méthode d'identification des techniques d'encodage et de toutes autres informations nécessaires à la communication multilingue entre clients et serveurs.

Ces méthodes ont toujours existé bien que la limite de leur usage ait toujours constitué un handicap majeur. Les utilisateurs continuent à s'échanger des données sans donner beaucoup de considérations aux jeux des caractères (normes ISO, Unicode, etc.) ni aux techniques d'encodage (ASCII, Binhex, Base64, QP, UTF-7, UTF-8, etc.) qu'ils utilisent. Ceci engendre énormément des distorsions (voire même des pertes) au niveau des données envoyées et reçues. La solution était de mettre en place un mécanisme d'encodage automatique que les agents utilisateurs « conçoivent » dans un format normalisé et interprété de manière uniforme par les machines communicantes. Un tel mécanisme aurait la tâche de reconnaître à la fois le jeu de caractères codés utilisé, le type de données concerné (Media type) et la technique d'encodage appliquée.

Parmi les solutions déjà entrées en vigueur, le standard MIME « *Multipurpose Internet Mail Extensions* », le concept des métadonnées (metadata), les feuilles de style et les serveurs intermédiaires (de conversion et de glyphes), jouent un rôle important dans la mise en place de mécanismes capables de transgresser l'unicité linguistique d'Internet.

Le standard MIME, créé en juin 1992, constitue un nouveau standard de messagerie Internet dont l'apport va au delà de la simple messagerie et le protocole SMTP pour couvrir également les documents WWW et leur protocole de transfert HTTP. La technique MIME consiste, par comparaison aux anciens systèmes de messagerie, à restructurer la zone d'en-tête où elle introduit de nouveaux champs pour la description de nouveaux types de contenus et d'organisation des messages.

Sur un plan linguistique, le courrier électronique était exclusivement limité au texte ASCII codé sur 7 bits. Pour coder un texte écrit dans des langues autre que l'anglais (langues latines et non latines), il était nécessaire de passer à une technique de codage sur 8 bits. C'était l'adoption de la solution ISO/IEC 8859-x (ISO 8859-1 ou ISO-Latin-1 pour les langues latines). Le passage à UCS-2 (Unicode) en substitut de ISO/IEC 8859-1 comme jeu de caractères codés standard est encore en phase de finalisation. MIME en a tiré profit pour permettre d'utiliser des caractères étendus codés sur 8 bits et rendre ainsi possible la communication et la transmission des données par messagerie électronique dans une grande multitude de langues. D'ailleurs, selon les recommandations du Consortium W3, le passage à UTF-8 à partir du 1 janvier 1999 sera un critère d'actualisation des agents utilisateurs de transport et de messagerie. Dans son rapport sur la messagerie Internet daté du mois d'août 1998, il est spécifié que « *any program created or revised after January 1, 1999, that cannot create mail using the UTF-8 charset should be considered deficient and lacking in standard internationalization capabilities* »^[112].

Parmi les nouveaux champs introduits par MIME afin de réaliser l'intégration des jeux des caractères codés sur 8 bits :

- Un champ d'en-tête "MIME Version" qui spécifie la version MIME utilisé par l'agent de messagerie.

[112] Internet Mail Consortium. Using International Characters in Internet Mail. Internet Mail Consortium Report : MAIL-I18N IMCR-010, 1 Aout, 1998. <http://www.imc.org/mail-i18n.html>

- ❑ Un champ d'en-tête « Content-Type » qui indique la catégorie (et la sous catégorie) des données dans le corps du message et renseigne sur la méthode d'encodage utilisé pour les mêmes données. Les différentes valeurs du champ « Content-Type » peuvent être « Text », « Multipart », « Application », « Message », « Image », « Audio », « Video » ...
- ❑ Un champ d'en-tête « Content-Transfer-Encoding » qui spécifie comment les données sont codées.
- ❑ Deux champs d'en-tête optionnels « Content-ID » et « Content-Description » pour une description particulière des données dans un corps de message.

Parmi ces quatre nouveaux champs, le deuxième (Content-Type) et le troisième (Content-Transfer-Encoding) constituent sans doute les deux champs clé du traitement et de l'échange des messages multilingues.

4.2.2.1.1 La paire « Content-Type » et « Charset »

Pour spécifier le mode d'encodage au moment du transfert des données sur les réseaux par le moyen des protocoles SMTP et HTTP, une structure appelée "Internet Media Type" ou "Mime Media Type" combine, comme il est défini par le RFC 2046^[113], les deux valeurs des champs "Content-Type" et "Charset" conformément à l'exemple suivant :

Content-Type: text/plain; charset=ISO 8859-6

Cette indication d'en-tête indiquera au système de transfert que la catégorie de données (Media Type) est du texte, que sa sous catégorie est du texte brut et que la technique de codage du texte est basée sur la norme ISO 8859-6 pour la langue arabe.

Le champ "**Content-Type**" est un exemple typique d'information qui devrait être généré automatiquement par les logiciels qui transmettent les données (i.e. logiciels de messagerie, serveurs Web) ou interprété par les logiciels qui les reçoivent (i.e. logiciels de messagerie, serveurs Web). Dans le cas des messages électroniques, les en-têtes sont positionnés au début du corps du message et le choix de les afficher ou de les cacher reste tributaire du choix de l'utilisateur (Idem pour les documents HTML via l'affichage des codes source).

Les exemples suivants illustrent l'usage de ces champs d'identification des messages conformément au paramétrage des options de l'agent de messagerie ou des choix de l'utilisateur.

```

Date: Mon, 17 Aug 1998 21:33:45 +0900
To: Mokhtar BEN HENDA <mokhtar.benhenda@isd.rnu.tn>
From: "Martin J. Duerst" <duerst@w3.org>
Subject: Re: i18n from Tunis
Cc: Tan Tin Wee <tinwee@pobox.org.sg>
In-Reply-To: <01bca7ec$8704f220$LocalHost@benmokht>
MIME-Version: 1.0
Content-Type: text/enriched; charset="ISO-2022-JP"
Content-Transfer-Encoding: quoted-printable
X-MIME-Autoconverted: from 8bit to quoted-printable by
tounes.gw.tn id OAA00595

Hello Mokhtar,
```

Figure 53 : En-tête d'un message codé en ISO-2022-JP transmis en QP

[113] <http://ds.internic.net/rfc/rfc2046.txt>


```

From: Roland Ducasse <ducasse@montaigne.u-bordeaux.fr>
To: Mokhtar BEN HENDA <Mokhtar.Benhenda@isd.rnu.tn>
Subject: UVF
Date: Mon, 31 Aug 1998 16:47:26 +0200
      .0
Content-Type: multipart/mixed; boundary="----
=_NextPart_000_01BDD4FF.09B44500"

-----=_NextPart_000_01BDD4FF.09B44500
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

Salut Mokhtar,

-----=_NextPart_000_01BDD4FF.09B44500
Content-Type: application/msword; name="UVFdef.rtf"
Content-Transfer-Encoding: base64

e1xydGYxXGFuc2

```

Figure 54 : En-tête d'un message codé en Multipart transmis en QP et Base 64

La figure suivante illustre les choix offerts à l'utilisateur par l'interface de messagerie sous Outlook Express.

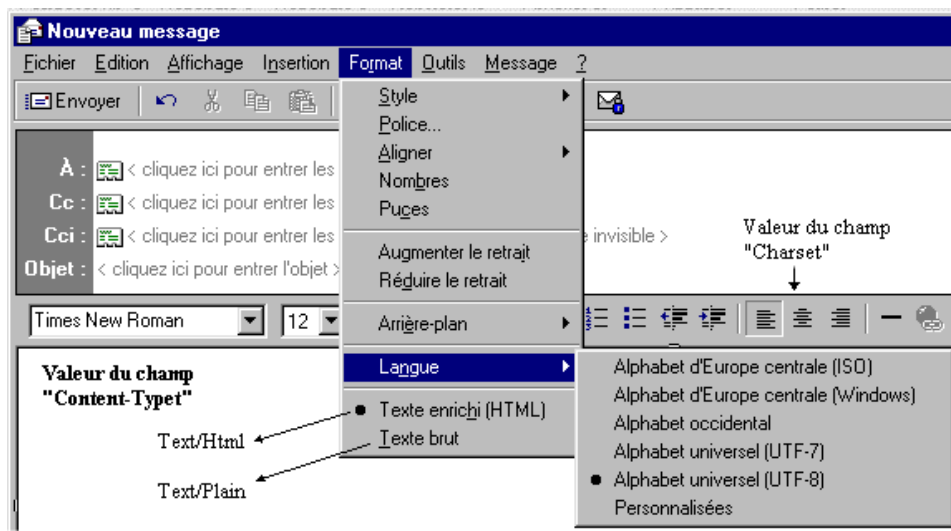


Figure 55 : Paramétrage des choix utilisateurs sous Outlook Express

4.2.2.1.2 Le champ « Content-Transfer-Encoding »

Le champ « Content-Transfert-Encoding » est un champ réservé à la messagerie électronique. Il précise dans une syntaxe particulière, une technique de transformation des caractères codés vers un schéma d'encodage capable de faciliter leur transmission par des protocoles spécifiques. Selon le standard HTTP 1.1 (voir RFC 2080, rubrique 14.12), l'entité du fichier en-tête « Content-encoding » est utilisé comme identificateur du type de document (media Type) concerné. Il indique la technique d'encodage et de compression (s'il y a lieu) utilisée et suggère par conséquent le mécanisme de décodage qu'il faudrait activer au moment de la réception du document.

Les techniques de transformation les plus référencées dans le champ « Content-Transfert-Encoding » sont :

- ❑ 7BIT
- ❑ BINARY
- ❑ 8BIT
- ❑ QUOTED-PRINTABLE
- ❑ BASE64

Les deux dernières techniques sont les plus couramment utilisées dans les opérations de transfert des données multilingues ou codés en binaires. Leur rôle est de permettre la transmission des valeurs 8 bits sur des canaux de transmission limités à 7 bits.

Quoted Printable

« L'imprimable Guillemeté » (!) comme on a parfois tendance à l'identifier, est destiné à transformer tous les caractères non ASCII en séries de caractères codés sur plus de 7 bits. Le mécanisme de fonctionnement de QP consiste à convertir toutes chaînes d'octets en chaînes similaires mais correspondant à des caractères ASCII. C'est en quelque sorte le même mécanisme que le format « %HH » que nous avons exposé ci-haut concernant l'encodage des valeurs Unicode en UTF-8. QP garde la correspondance normale entre les caractères ASCII et leurs équivalents en chaînes d'octets, mais procède à une conversion des octets au delà des 128 premières positions selon le principe suivant :

Un Octet « n » est représenté comme séquence de trois octets différents qui correspondent aux valeurs ASCII du signe égal (=) suivi des deux chiffres de la valeur octale de « n ». Si l'encodage QP est appliqué à des séquences d'octets qui représentent des caractères codés selon ISO/IEC 8859-6, les caractères ASCII sont rendus tels quels alors que les caractères arabes sont rendus en fonction de leurs équivalents en valeurs octales précédées du signe « = ».

Puisque QP est un mécanisme de simple conversion d'une chaîne d'octets à une autre, il est considéré comme un codage sans risques et pourrait donc être appliqué à toutes sortes de caractères destinés à être transmis dans toutes les langues. Il demeure cependant nécessaire que l'encodage QP soit traité par un programme capable de reconvertir les codes vers un format lisible par l'être humain (présence des polices de caractères appropriées).

Dans l'exemple suivant d'un message rédigé en arabe et codé en QP, les caractères arabes sont codés par un mécanisme octal que l'agent utilisateur a pu reconvertir dans une forme lisible. (N.B. : la figure représente un message arabe composé d'une chaîne invertie de caractères car démontrée sous la version latine de MS-Outlook Express qui ne gère pas la bidirectionnalité au niveau du corps du message. La justification droite est forcée par le bouton d'alignement droite).

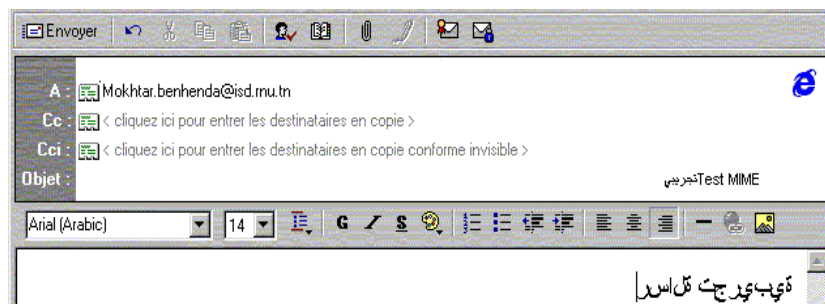


Figure 56 : Rendu d'un message arabe codé en QP

```

...
MIME-Version: 1.0
Content-Type: multipart/alternative;
    boundary="-----_NextPart_000_0007_01BDDFE5.7A2B3A00"
X-Priority: 3
X-MSMail-Priority: Normal
X-Mailer: Microsoft Outlook Express 4.71.1712.3
X-MimeOLE: Produced By Microsoft MimeOLE V4.71.1712.3

Message en plusieurs parties et au format MIME.

-----_NextPart_000_0007_01BDDFE5.7A2B3A00
Content-Type: text/plain;
    charset="utf-8"
Content-Transfer-Encoding: quoted-printable

=D8=B1=D8=B3=D8=A7=D9=84=D8=A9 =
=D8=AA=D8=AC=D8=B1=D9=8A=D8=A8=D9=8A=D8=A9

-----_NextPart_000_0007_01BDDFE5.7A2B3A00
Content-Type: text/html;
    charset="utf-8"
Content-Transfer-Encoding: quoted-printable

<!DOCTYPE HTML PUBLIC "-//W3C//DTD W3 HTML//EN">
<HTML>
<HEAD>

<META content=3D"text/html; charset=3Diso-8859-6" =
http-equiv=3DContent-Type>
<META content=3D'"MSHTML 4.71.1712.3"' name=3DGENERATOR>
</HEAD>
<BODY bgColor=3D#ffffff>

```

Figure 57 : Source du message arabe codé en QP

Base 64

La seconde technique d'encodage « Base 64 » est plutôt utilisée pour le transfert des fichiers binaires et les textes codés sur plus de 7 bits transmis sous forme de pièces rattachées ou incorporées dans les messages. Contrairement à BinHex ou uuencode, la technique Base64 (RFC 2045, section 6.8) est standardisée. Elle résiste mieux aux « déformations » durant le transport des données sur le réseau Internet.

L'exemple suivant illustre les deux modes de codage QP et Base 64 dans le même corps de message.

```

...
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="-----
=_NextPart_000_01BDD4FF.09B44500"

-----=_NextPart_000_01BDD4FF.09B44500
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

```

```

Salut Mokhtar, je ... dont je vous ai parl=E9. J'attends (?) de
vos nouvelles.
A+
----- =_NextPart_000_01BDD4FF.09B44500
Content-Type: application/msword; name="UVFdef.rtf"
Content-Transfer-Encoding: base64

e1xydGYxXGFuc2lcYW5zaWNwZzEyNTJcdWMxIFxkZWZmMFxkZWZsYW5nMTAzM1xkZW
ZsYW5nZmUxMDM2e1xmb250dGJse1xmMFxmcm9tYW5cZmNoYXJzZXQwXGZwcnEye1wq
XHBhbm9zZSAwMjAyMDYwMzA1MDQwNTAyMDMwNH1UaW1lcYB0ZXcgUm9tYW47fXtcZj
FcZnN3aXNzXGZjaGFyc2V0MFxmcm9tYW5cZmNoYXJzZXQwXGZwcnEye1wq
MjAyMDR9QXJpYWw7fQ0Ke1xmM1xmcm9tYW5cZmNo

```

Figure 58 : Les deux modes d'encodage QP et Base 64 dans un message Multipart

Si l'on observe le contenu d'un fichier codé en Base64, on notera qu'à l'instar de Uuencode^[114] et de Binhex^[115], tout le contenu est codé dans une suite de chiffres et de lettres ASCII (65 en tous et 76 caractères par ligne = A-Za-z0-9+/=) afin de garantir leur transfert sans aucun risque de distorsion à travers les routeurs et les passerelles Internet. Le fichier codé en Base64 augmente généralement de taille d'un ordre de 33% par rapport à son original.

L'avantage de ces deux techniques de codage MIME est double : d'une part ils ne sont pas propriétaires contrairement à Binhex (plate-forme Macintosh) ; et d'autre part ils sont universels contrairement à Uuencode (plate-forme Unix) et ses nombreuses variantes.

Par contre, elles ne sont pas les deux seules alternatives de codage que MIME a introduites. Selon Guy Brand, « *Un logiciel supportant MIME est capable d'interpréter des messages dont le contenu n'a pas été codé, comme ceux véhiculés en "7 bit", "binary" ou "8 bit"* »^[116]. Il précise que ces deux derniers, peuvent subir des distorsions au delà du 8^e bit par les dispositifs des réseaux via lesquels ils transitent.

En définitive, la conformance à MIME est selon Brand « *un label garantissant que le logiciel est capable de garantir l'interopérabilité des messages MIME* »^[117]. Quand un agent utilisateur ou un logiciel client s'identifie aux cinq fonctionnalités suivantes, il est dit "conforme à MIME" :

1. Insertion d'un champ d'en-tête "MIME-Version: 1.0" dans tous les messages émis ;
2. Reconnaissance du champ "Content-Transfer-Encoding" et déchiffrement des données codées avec QP ou base64 ;
3. Reconnaissance et interprétation du champ d'en-tête "Content-Type" pour éviter d'afficher des données non textuelles identifiées par des paramètres autres que "text" ;
4. Traitement direct des types de données suivants : Texte (Text), Image, audio et vidéo, Application, Multipart et Message ;
5. Traitement de tout message ayant un champ « Content-Type » nul ou inconnu (i.e. "application/octet-stream").

[114] Unix to Unix Encoding : <http://kb.indiana.edu/data/acup.html?cust=13752>

[115] Binary hexadecimal : <http://kb.indiana.edu/data/acma.html?cust=13752>

[116] Guy BRAND. Qu'est-ce que MIME ? 17 janvier 1995 (Révisé le 4 juin 1997). <http://wwwchimie.u-strasbg.fr/membres/GB/MIME.html>

[117] Guy BRAND. Qu'est-ce que MIME ? Op.Cit.

L'interopérabilité pourrait être ainsi garantie entre plusieurs agents utilisateurs. Deux précautions supplémentaires sont quand même à prendre en considération :

- pour que deux correspondants puissent s'affranchir des déformations des caractères engendrées par les protocoles de transmission, des logiciels conformes à MIME doivent être utilisés par les deux correspondants aux deux bouts du canal de communication.

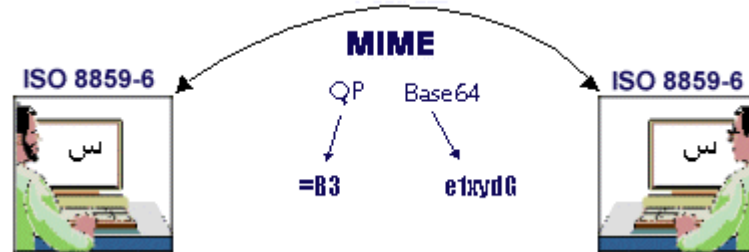


Figure 59 : Échange codé en MIME entre jeux de caractères identiques

- tirant profit de l'avantage MIME d'utiliser des jeux des caractères étendus, les deux correspondants doivent installer des polices de caractère capables d'afficher sur écran les messages qu'ils se transmettent.

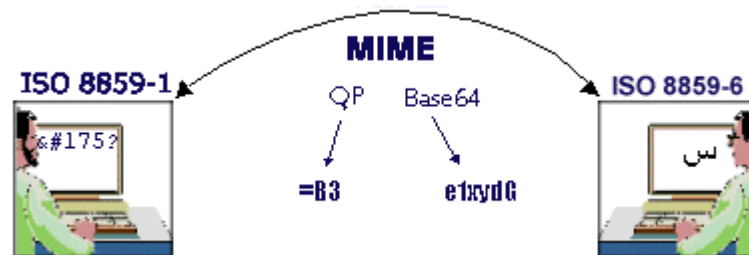


Figure 60 : Échange codé en MIME entre jeux de caractères différents

Les paramètres « Content-Type » et « Content-Transfer-Encoding » sont entièrement paramétrables par l'utilisateur sur certaines interfaces Homme-Machine. L'exemple suivant de Microsoft Outlook Express est un exemple **non** généralisé de la convivialité de l'interface utilisateurs pour la définition d'un choix des techniques de codage et de transmission des données.

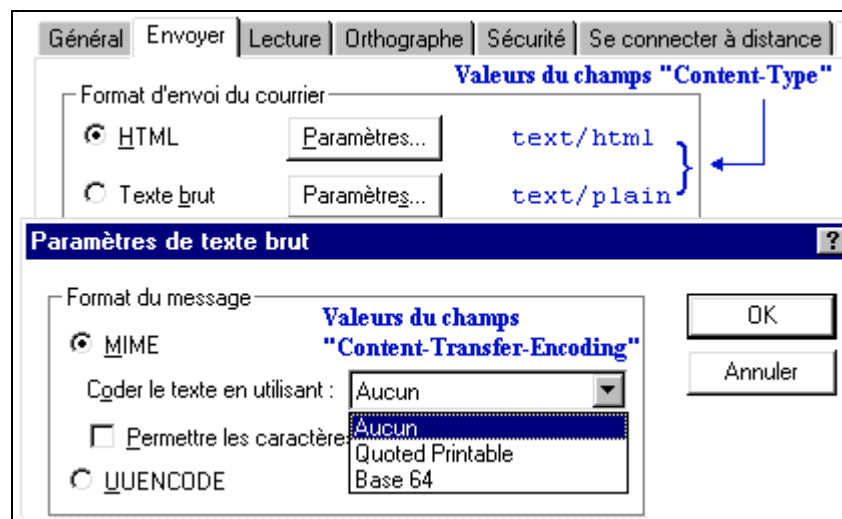


Figure 61 : Paramétrage utilisateur des modes Texte et des techniques de codage sous Outlook Express

Cet autre exemple illustre les options de l'agent de messagerie *Pegasus Mail for Windows* pour le codage des messages et des fichiers qui leur sont rattachés.

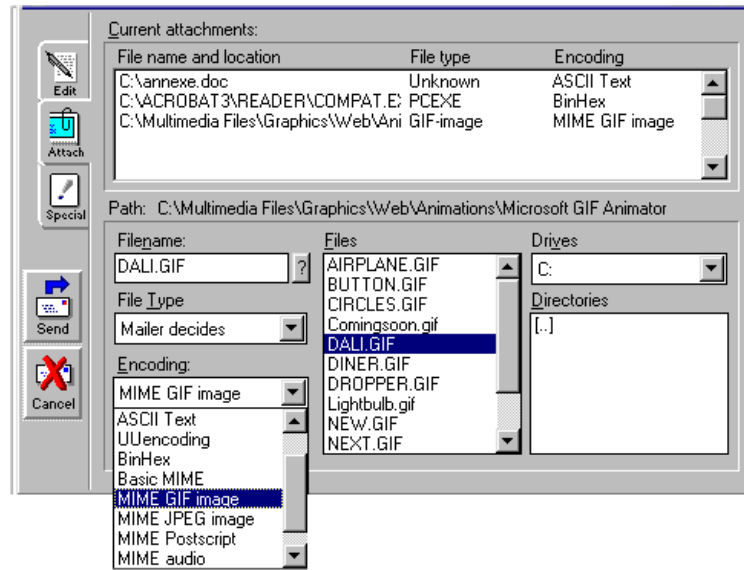


Figure 62 : Options de codage sous Pegasus Mail

4.2.2.2 HTTP/MIME dans un environnement WWW : étiquetage et négociation multilingues

À l'instar de l'identification du jeu de caractères utilisé à travers l'en-tête HTTP/MIME d'un message codé en Text/* (ou * peut être un type de message structuré « HTTP » ou en texte brut « Plain »), les documents circulant sur le Web font également recours aux champs et attributs MIME pour la résolutions des versions linguistiques et l'identification des jeux des caractères utilisés dans le définition de leur formes internes et leurs transferts sur le Web.

À priori, deux situations sont adressées par la désignation des attributs linguistiques des ressources WWW :

1. l'étiquetage linguistique de documents ou des parties des documents Web ;
2. la négociation des versions linguistiques au moment du transfert des documents Web.

4.2.2.2.1 L'étiquetage multilingue des documents WWW

L'identification des jeux des caractères au moment de la conception d'une ressource WWW applique généralement la méthode d'étiquetage de MIME que nous avons exposée dans la partie précédente avec les deux attributs les plus significatifs « CONTENT » et « CHARSET » sous la forme :

Content-Type: */*; charset=jeu de caractères

Exemple :

Content-Type: text/plain; charset=ISO 8859-1

Cette méthode a toujours constitué le modèle d'étiquetage standard des langues de l'Europe de l'Ouest basé sur le jeu des caractères ISO-8859-1 adopté par défaut pour les documents HTML. Pour afficher des documents dans des langues autres que latines (i.e. arabe), il était nécessaire de choisir le jeu de caractères alternatif qu'il fallait indiquer comme valeur du champ « CHARSET » de la manière suivante :

Content-Type: text/plain; charset=ISO 8859-6

Cette manière, selon les affirmations de Andrew Daniel (Université de Vancouver)^[118], causait des erreurs d'interprétation de la part de certains navigateurs anciens qui l'interpréteraient comme un fichier binaire pour lequel ils lançaient des opérations de sauvegarde sur disque au lieu d'afficher la page dans le jeu de caractères spécifié. Cette fonctionnalité d'affichage était alors rectifiée en utilisant les paramètres « META » du langage HTML conformément à la structure suivante :

<META HTTP-EQUIV= « Content-type » CONTENT= « text/html ; charset=iso-8859-6 »

L'introduction de cette nouvelle forme d'identification des valeurs linguistiques et des techniques de codage des documents à travers l'élément « META » a permis aux navigateurs de changer automatiquement de fontes chaque fois qu'il y a un nouvel appel pour un jeu de caractères codés différent.

La zone des métadonnées (METADATA) est imbriquée dans le même en-tête d'un document HTML pour fournir davantage de données, entre autres linguistiques, sur le document transmis conformément aux spécificités MIME.

Les éléments « META » reprennent intégralement les valeurs de l'attribut « CONTENT-TYPE » avec ses deux paramètres « Content » et « charset » qui identifient respectivement le type de document et le codage qui lui est appliqué.

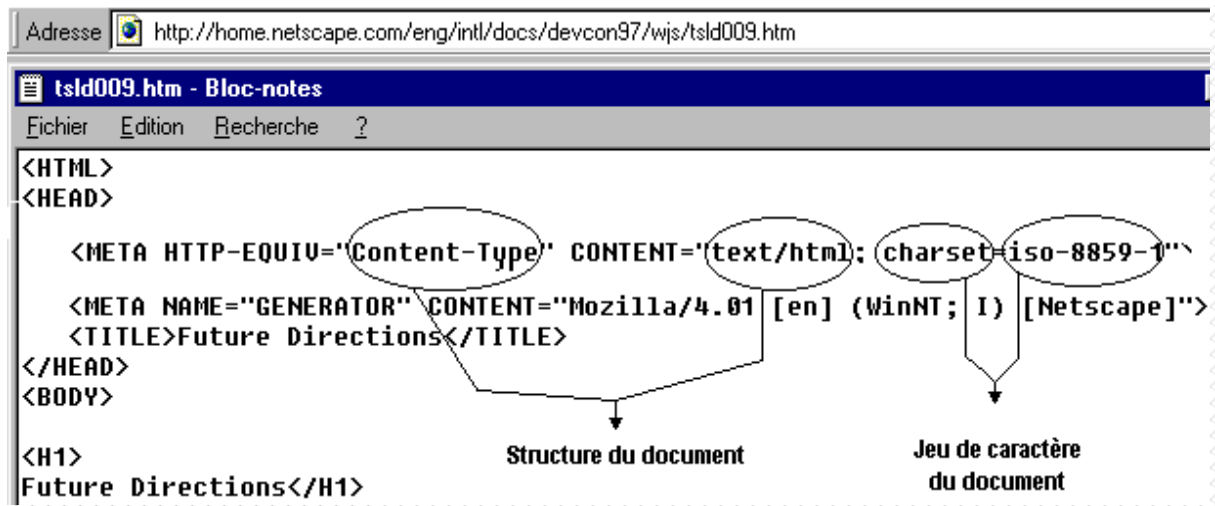


Figure 63 : Étiquetage linguistique simple par zone METADATA

Le problème inhérent à cette méthode est le fait que beaucoup de serveurs ignoraient cette précaution même quand le jeu de caractères utilisé était différent des jeux de caractères ASCII et Latin-1 qui étaient pour longtemps des jeux par défaut. Pour remédier à cet oubli, HTML permet

[118] Andrew DANIEL. How to make a Multilingual Web server. <http://vancouver-webpages.com/multilingual/howto.html>

aux auteurs des documents HTML de spécifier eux même cette particularité à travers l'élément META indiqué dans l'en-tête du document.

Devant la situation chaotique dans l'attribution des éléments META de l'en-tête des documents HTML, des tentatives de normalisation commencent à se faire entendre pour recommander et harmoniser l'usage des étiquettes META selon des principes uniformes et généralisés. Dublin Core^[119] en est un exemple très avancé dans la communauté des spécialistes des bibliothèques numériques et des documents structurés qui vise depuis sa création en mars 1995 à Dublin (Ohio-USA) à adresser le problème de la description unifiée des ressources d'information électroniques et donc de leur localisation dans un contexte réseaux. Des outils concepteurs de métadonnées pour les documents HTML ont été même mis au point pour aider l'uniformisation de cette partie stratégique dans la structure des documents électroniques.

L'exemple suivant illustre une procédure de création d'éléments META par l'outil META BUILDER de l'université de Vancouver^[120] mis en accès public pour la conception d'en-têtes standardisés conformes à Dublin Core. Nous reprenons dans la figure suivante la partie du formulaire relative à la saisie des variables concernant les jeux des caractères utilisés, en l'occurrence l'élément « CHARSET ».

Owner: Mokhtar.Benhenda@isd. (Email address)

Expires: 365 days e.g. +25 days or Sunday, 12-M

Charset (Recommended if not ISO8859-1)

Charset: ISO-8859-6 (other:)

Language: other (other: arabic)

Dialect (country): other (other: Tunisia)

Figure 64 : Extrait du formulaire de définition des éléments META par *META BUILDER*

Le texte suivant illustre le rendu du traitement du formulaire sous forme d'un en-tête META normalisé à inclure au début d'une page de sources HTML.

```
The beginning of your HTML document (root page if using frames) should
look like this (clip or save as TEXT):
<HTML LANG=ar-TN>
<HEAD PROFILE="http://purl.org/metadata/dublin_core">
<META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=ISO-8859-6">
<META HTTP-EQUIV="Expires" CONTENT="Mon, 20 Sep 1999 07:48:36 GMT">
<TITLE>charset iso 8859-6</TITLE>
<LINK REV=made href="mailto:Mokhtar.Benhenda@isd.rnu.tn">
<META NAME="description" CONTENT="Test de création d'un document html à
base d'un charset iso 8859-6">
<META NAME="rating" CONTENT="General">
<META NAME="Vw96.objecttype" CONTENT="Document">
```

[119] Stuart WERIBEL, Jean GODBY, Eric MILLER, Ron DANIEL. OCLC/NCSA Metadata Workshop Report. 1995.
http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html

[120] META BUILDER : <http://vancouver-webpages.com/META/mk-metas.html>

```

<LINK REL=SCHEMA.VW96 HREF="http://vancouver-webpages.com/META/VW96-
schema.html">
<META NAME="ROBOTS" CONTENT="ALL">
<META NAME="DC.Title" CONTENT="charset iso 8859-6">
<META NAME="DC.Creator" CONTENT="Meta builder">
<META NAME="DC.Subject" CONTENT="iso 8859-6 langue arabe négociation i18n
html meta">
<META NAME="DC.Description" CONTENT="Test de création d'un document html à
base d'un charset iso 8859-6">
<META NAME="DC.Language" SCHEME="RFC1766" CONTENT="AR">
<!-- Metadata generated by http://vancouver-webpages.com/META/mk-metas.html -
->
</HEAD>
Note: Certain HTML editors mangle META elements; you may wish to use a
simple editor like Notepad or vi to add these.

You may generate the following HTTP headers using a CGI script or
otherwise (e.g. Apache .meta file) in your server:
Content-language: ar-TN
Expires: Mon, 20 Sep 1999 07:48:36 GMT
Content-Type: text/html; charset=ISO-8859-6
You might wish to include a charset attribute in the anchor in the parent
document, e.g.
<A HREF="parent.html" CHARSET="ISO-8859-6">charset iso 8859-6</A>

```

Dans l'attente d'arriver à ce stade d'harmonie et de consensus généralisé, l'identification des éléments META des documents HTML continue à se faire de façon sporadique et aléatoire.

Il est toutefois de rappeler que la déclaration META doit être utilisée uniquement quand l'encodage des caractères est organisé de façon à ce que les caractères soient interprétés en ASCII au moins jusqu'à ce que l'élément META soit atteint. Les agents utilisateurs pourront alors interpréter correctement l'élément META conformément à la spécification linguistique de l'élément « Charset ». Ceci nous oblige à soulever la problématique quasiment ignorée de l'i18n des littéraux de formatage (éléments en-têtes, attributs, balises HTML, etc.) exclusivement limités au jeu de caractères ASCII. Cette problématique constituerait à notre sens une continuité de l'i18n des URI que nous avons soulevée dans les paragraphes précédents. D'autres éléments d'analyse seront toutefois fournis dans les paragraphes suivants relatifs aux apports d'Unicode et les références numériques des caractères dans l'i18n des zones systèmes.

La figure suivante illustre l'usage des champs MIME dans une zone METADATA pour l'identification des attributs linguistiques d'un document HTML rédigé en langue arabe.

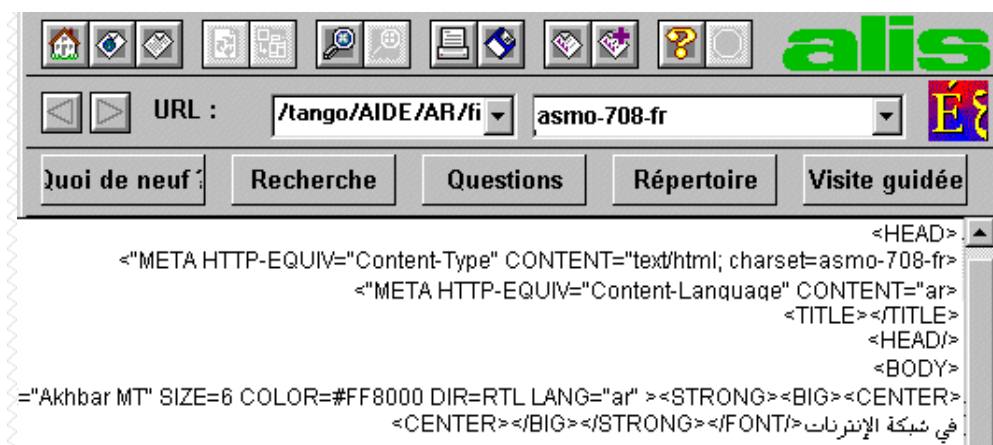


Figure 65 : Etiquetage linguistique d'un document HTML arabe par l'élément « META » sous Tango

Dans cet exemple, la zone META décline l'identité linguistique du document par les deux rubriques « META HTTP-EQUIV » suivantes :

- ❑ **Content-Type** "CONTENT="text/html; charset=asmo-708-fr : ceci indique qu'il s'agit d'un document textuel en format HTML codé selon le jeu de caractères codés ASMO 708 ;
- ❑ **Content-Language** "CONTENT="ar : ceci précise que la valeur culturelle du document est explicitement relative à la langue arabe sachant que la norme ASMO 708 (§ III-3.3) pourrait être utilisée pour d'autres langues utilisant l'écriture arabe. L'entité du champ « Content-Language » décrit la/les langue(s) naturelle(s) du public cible auquel le document est initialement destiné.

Ces mesures sont, à notre sens, d'un penchant plutôt statique. C'est à dire que la désignation ou l'étiquetage de la version linguistique d'un document s'accomplit à la source par l'auteur ou le serveur hébergeant la ressource sans considération des variables linguistiques des utilisateurs virtuels. Il s'agit d'une construction d'un document à la base des spécificités particulières du serveur hébergeur ou de l'auteur concepteur du document en question. La situation devient plus critique quand le document est soumis à la diffusion par navigation hypertextuelle (liens), par adressage ciblé (URI) ou par formulaires de requêtes (moteurs de recherche) provenant de la part d'utilisateurs virtuels à profils linguistiques très divers. Le mécanisme de la négociation linguistique entre alors en considération pour entamer plusieurs formes de transactions entre clients et serveurs à la fois.

4.2.2.2 La négociation des attributs linguistiques des ressources WWW

La négociation linguistique entre client et serveur pour l'échange des documents multilingues fait actuellement partie du modèle générique et complexe de la négociation des choix de traitement des données (encodage des données, format des fichiers ...). L'acquis en la matière est que les systèmes actuels sont parvenus à rattacher toutes les variantes linguistiques d'un document unique à une seule adresse électronique ou à un seul processus de requête évitant à l'utilisateur final de rechercher des pointeurs éparpillés sur l'interface pour atteindre des versions linguistiques multiples (Ce n'est pourtant pas toujours évident si l'interface et les liens sont dans une langue inconnue pour l'utilisateur). L'objectif final de la négociation est, comme l'a exprimé Anoosh Hosseini dans une intervention sur une liste de diffusion *"So what we are talking about is a generic mechanism to negotiate what is sent down by the server. Now the client may be out of date, but if the servers are supporting say CERN's .multi feature, I as the server would know that the client does not support language XX and would send down a Latin1 doc explaining why the user cannot read this document. This is the proper response rather than saving unreadable material to disk"*^[121].

Le principe de la négociation est conçu selon le protocole HTTP comme une relation Question/Réponse entre client et serveur. Un client transmet une requête à un serveur sous forme d'un formulaire ou d'une adresse URI qui seront suivis implicitement par un message invisible à l'utilisateur contenant des indications sur les paramètres de la requête, l'agent utilisateur etc. Le serveur répond avec une série d'information, également invisible à l'utilisateur, dans laquelle il inclut des valeurs relatives aux versions du protocole utilisé et du système d'exploitation du serveur, du nom de l'interface et du programme CGI employés, etc.

[121] Anoosh HOSSEINI (anoosh@gorgan.mti.sgi.com). Wed, 8 Feb 95 21:07:12 EST

Le modèle suivant, obtenu sur le serveur « *vancouver-webpages.com* » résume les types d'information que clients et serveurs émettent durant une requête d'information.

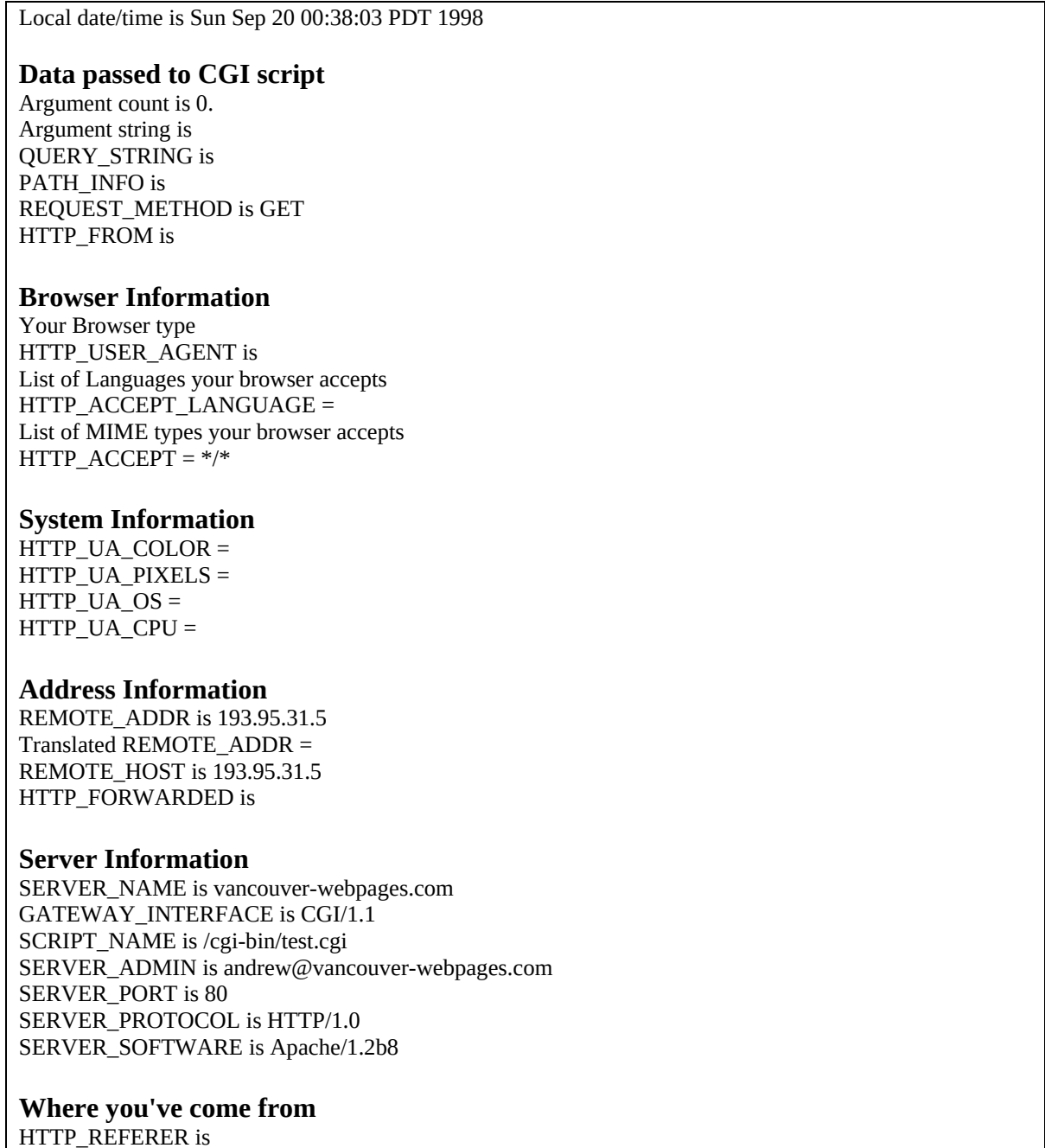


Figure 66 : Modèle de messages entre clients et serveurs pour traiter une requête par formulaire

L'objectif primordial de cette négociation entre client et serveur est de parvenir à établir la « meilleure » réponse à une requête utilisateur. Or, la « meilleure » réponse est toujours relative aux spécificités particulières de chaque utilisateur. Pour cette raison, HTTP prévoit une possibilité de négociation du contenu des documents entre clients et serveurs dans les cas où plusieurs possibilités de réponses sont disponibles à une requête quelconque.

Pour atteindre cette interopérabilité, il est primordial que l'agent utilisateur dispose d'un minimum d'indications sur les performances linguistiques du serveur en termes des techniques d'encodage qu'il est capable d'interpréter pour résoudre l'équation de recherche qui lui est

transmise par formulaire. Le serveur devrait également fournir une réponse (le cas où il en dispose de plusieurs versions linguistiques) qui correspond au profil de l'utilisateur sinon qui s'en rapproche le plus.

En réalité HTTP permet d'établir deux sortes de négociations : l'une conduite par le serveur au moment du traitement de la requête et l'autre accomplie par le client au moment de l'émission de la requête.

Négociation par serveur

Selon le standard HTTP 1.1, si la sélection du « meilleur » document devait se faire au niveau du serveur, un algorithme spécialisé (installé sur le serveur) effectuerait la sélection appropriée parmi une variété de réponses disponibles, en fonction d'un ensemble de critères vérifiés sur les réponses elles mêmes (i.e. langue, technique de codage ...) ou le contenu de certains champs d'en-tête de la requête.

La négociation par serveur est avantageuse dans les deux cas de figure quand l'algorithme de sélection de la « meilleure » réponse est difficile à accomplir par l'agent utilisateur ou quand le serveur effectue une analyse de pertinence (heuristique) qui satisfera la requête de l'utilisateur du premier coup. La contribution de l'agent utilisateur dans l'analyse de pertinence par le serveur est importante à ce niveau car il pourrait à tout moment adjoindre à l'en-tête du formulaire de requête des champs qui identifieraient ses préférences dans la désignation de la réponse. HTTP 1.1 met à la portée des agents utilisateurs la série de champs de préférences linguistiques suivantes :

- ❑ Accept (HTTP 1.1 section 14.1),
- ❑ Accept-Charset (HTTP 1.1 section 14.2),
- ❑ Accept-Encoding (HTTP 1.1 section 14.3),
- ❑ Accept-Language (HTTP 1.1 section 14.4), et
- ❑ User-Agent (HTTP 1.1 section 14.42).

Le schéma suivant illustre la méthode par laquelle un agent utilisateur spécifie au serveur ses préférences linguistiques pour négocier la nature de la réponse souhaitée.

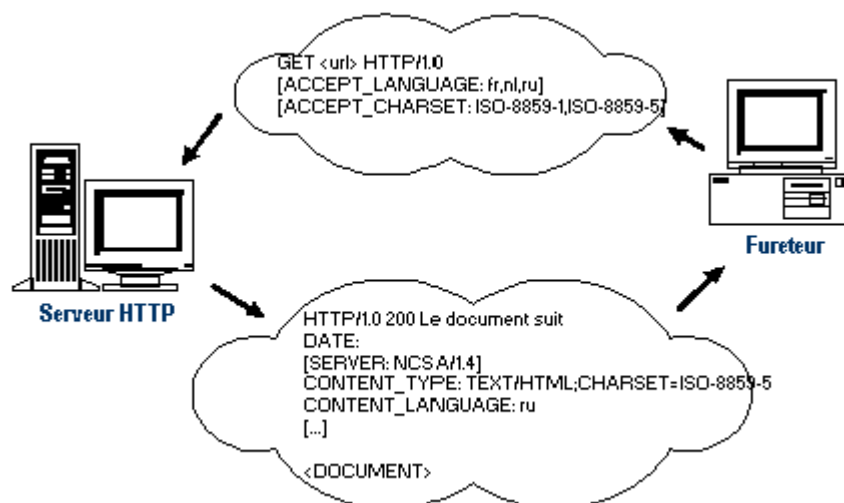


Figure 67 : Négociation de réponse entre client et serveur à base de critères linguistiques

Source : Jean Bourbonnais, François Yergeau, Language on Internet. http://www.crim.ca/inet96/papers/a5/a5_3.htm

Dans cet exemple, l'agent utilisateur a spécifié qu'il accepterait en réponse à une requête posée, des documents codés en ISO 8859-1 et ISO 8859-5 (exprimés par champ « ACCEPT_CHARSET ») dans un ordre prioritaire de préférences linguistiques spécifiées par le champ « ACCEPT_LANGUAGE ». Cet en-tête a été traité par le serveur HTTP qui a établi en conséquence le choix d'une des réponses disponibles qui s'accorde le « mieux » avec les préférences de l'utilisateur.

La réponse du serveur peut toutefois varier en fonction d'autres critères de la requête ou prise en dehors de la zone d'en-tête elle-même.

La négociation par serveur connaît par contre certaines lacunes dont les plus importantes seraient :

- ❑ l'impossibilité de déterminer de façon exacte ce qui pourrait être le « mieux » pour l'utilisateur car ceci nécessiterait une connaissance parfaite du profil de l'utilisateur et de ses intentions d'usage du document restitué ;
- ❑ demander à l'agent utilisateur de spécifier ses performances avec chaque requête s'avérerait inutile partant du fait que peu de réponses disposent de représentations variables susceptibles d'être soumises à des opérations de sélection de choix ;
- ❑ la limitation des zones Cache publiques qui pourraient utiliser les mêmes réponses pour des requêtes d'autres utilisateurs ;

Négociation par agent utilisateur

Dans ce genre de négociation linguistique, la « meilleure » représentation d'une réponse est accomplie par l'agent utilisateur après réception d'une réponse initiale du serveur d'origine. La sélection s'effectue à partir d'une liste d'alternatives de réponses disponibles que le serveur envoie avec l'en-tête de sa réponse initiale sous forme d'un champ « ALTERNATES » nouvellement introduit par HTTP 1.1 (section 19.6.2.1) comme démarche d'amélioration de la négociation client/serveur.

La négociation par agent utilisateur est avantageuse pour faire le choix parmi une liste de réponses variables pour la même requête (variabilité de type de document, de langue, de type de codage ...) qui ne porte aucune indication sur le profil utilisateur que le serveur d'origine aurait pu exploiter pour négocier à son niveau la « meilleure » réponse à transmettre.

La négociation par agent utilisateur souffre cependant d'une lourdeur de procédure inhérente à la retransmission de la même requête pour obtenir la « meilleure » alternative sélectionnée par l'agent utilisateur sur proposition préalable du serveur d'origine. Sachant que la spécification actuelle de HTTP ne définit aucun mécanisme de sélection automatique par l'agent utilisateur, il serait convenable de développer une pareille procédure comme extension à HTTP 1.1.

Dans la perspective de mieux gérer cette négociation linguistique de part et d'autre du processus communicationnel, Gavin Nicol recommande de porter certaines modifications au niveau des clients et des serveurs^[122].

[122] Gavin NICOL. The Multilingual World Wide Web. November 1994. <http://www.ebt.com:8080/docs/multilingual-www.html>

- ❑ Tous les clients doivent envoyer un paramètre « Accept-Charset » si le jeu de caractères des documents à restituer est autre que ISO-8859-1 ;
- ❑ Tous les clients doivent supporter les références des caractères numériques conformes à Unicode comme caractères de base ;
- ❑ Tous les clients doivent être capables d’afficher les caractères Unicode en format UTF-8 ou UCS-2 ;
- ❑ Tous les clients doivent adapter leurs méthodes de saisie et de rendu aux effets des « Locales » ;
- ❑ Les serveurs doivent à leur tour libeller les données transmises par champ « CONTENT-TYPE ». Le paramètre « charset » doit supporter autres techniques de codage que US-ASCII ou ISO 8859-1 ;
- ❑ Les serveurs doivent accepter le type de données libellées « multipart/form-data » ;
- ❑ Les serveurs doivent permettre la conversion vers Unicode ;
- ❑ Des serveurs intermédiaires de conversion doivent être mis en place ;
- ❑ Des serveurs de glyphes doivent également être prévus pour permettre un rendu Unicode intégral sans besoin de fontes appropriées.

Quoi qu’il en soit, la situation actuelle est encore en phase embryonnaire malgré les évolutions que ne cessent d’enregistrer les efforts de l’i18n d’Internet. La communauté internationale continue à faire face à une accentuation des problèmes de déchiffrement, de traitement et d’affichage des jeux des caractères utilisés dans les systèmes ouverts et distribués. Dans l’attente de la solution miracle, celle de l’i18n et du multilinguisme (nuance !)^[123] en l’occurrence la transparence linguistique grâce à l’adoption unanime d’Unicode et d’ISO 10646, les solutions intermédiaires continuent à émerger de manière sporadique pour traiter l’immédiat par des propositions diverses qui varient entre l’imagerie des textes, des glyphes et caractères, l’enrichissement du balisage HTML, l’adjonction des services de traduction et de traductique, le décodage intermédiaire par services interposés ...

4.2.2.3 Alternatives d’i18n des transactions HTTP

En effet, s’il est évident que l’évolution de HTTP et le gain de confiance que le codage sur plusieurs Octets est en train d’enregistrer conduiraient inéluctablement vers une solution unique et univoque pour résoudre cette polémique du choix des jeux des caractères non latins, cette perspective n’est pas prévue pour des échéances à courts termes. La solution Unicode a certainement permis de résoudre relativement cette charge linguistique par la réduction du grand nombre de jeux des caractères codés à un seul (du moins pour des perspectives encore à longs termes). Pour traiter l’immédiat, les solutions ad hocs sont actuellement proposées et étudiées

[123] L’i18n des services d’information est le fait d’adapter un outil à une langue particulière (i.e. français OU arabe). Le Multilinguisme est le fait de rendre cet outil capable de traiter convenablement deux langues ou plus au même temps (i.e. français ET arabe). L’i18n des ressources veut ainsi dire que l’utilisateur est conscient de l’existence d’autres versions d’un même document. Le multilinguisme implique aussi que les navigateurs deviennent capables de traiter toutes les langues à la fois.

pour anticiper à cette transition future vers le « tout Unicode ». Nous commenterons ci-après quelques-unes de ces propositions et nous exposerons notre propre initiative pour la transparence linguistique des documents sur le Web par une proposition non propriétaire.

4.2.2.3.1 Serveurs intermédiaires

L'une des propositions en guise d'alternative aux difficultés du transcodage, de la transmission et du traitement multilingue des documents Internet consisterait à mettre en place des serveurs intermédiaires qui joueront le rôle de relais entre un serveur dont les capacités de traitement linguistiques seraient non conformes à une requête d'un client quémandeur de ressources.

a. Serveurs de transcodage

Il est évident que parfois les serveurs ne sont pas capables de traduire convenablement les jeux des caractères et les modes d'encodage des documents qu'ils détiennent vers ceux demandés par les clients à travers leurs requêtes d'information. Dans ce cas de figure, le serveur redirige le client vers un serveur intermédiaire afin d'accomplir le transcodage nécessaire. Le serveur intermédiaire agirait ainsi comme serveur proxy pour le client tout en changeant quand nécessaire les jeux des caractères incompréhensibles du serveur vers un jeu approprié pour le client.

Bien que ce modèle, proposé initialement par Larry Masinter^[124], ne soit pas conforme aux recommandations d'usage des réseaux, il a l'avantage de ne pas imposer à tous les serveurs la fonction de transcodage des jeux des caractères et de reléguer cette fonction à des serveurs dédiés avec tout le gain de performance que cela engendrerait. L'avènement de la solution Unicode fera de ce modèle une proposition transitoire.

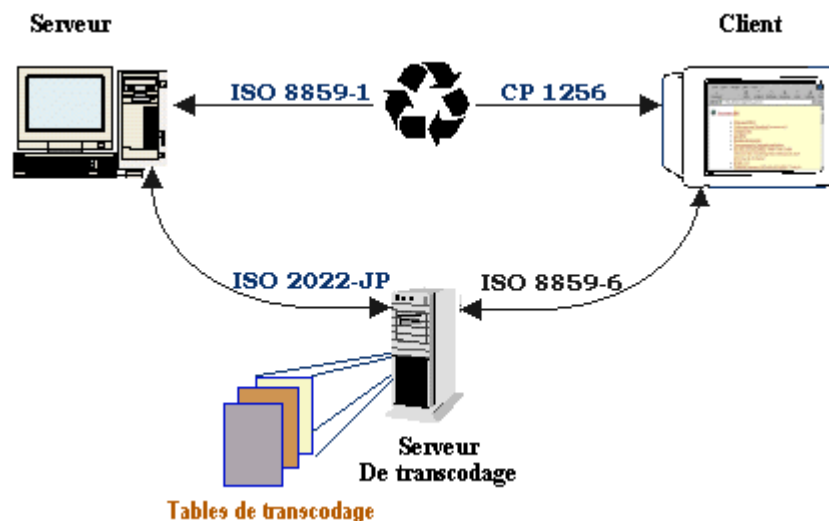


Figure 68 : Modèle de serveurs de transcodage intermédiaires

Dans l'exemple de la figure précédente, le jeu de caractères codés d'un client qui fonctionne en mode Windows CP 1256 (Bilingue arabe/latin) serait converti directement de et vers ISO 8859-1 alors que pour un document codé en ISO 2022-JP, un transcodage par serveur intermédiaire à eu lieu entre ISO 2022-JP et ISO 8859-6 suivi d'un transcodage direct entre ce dernier et CP 1256.

[124] Cité par Gavin NICOL. The Multilingual World Wide Web. Op.Cit.

b. Serveurs de glyphes

Pour la fonction d’affichage, des serveurs de glyphes sont également proposés pour jouer le rôle de proxy à des clients vers lesquels un serveur pourrait rediriger le résultat d’une requête afin de la transmettre au client dans un rendu compréhensible par l’utilisateur. L’exemple précédent aurait été complété par l’aiguillage sur un serveur de glyphes pour reproduire en format image les caractères ISO 2022-JP qui n’ont pas d’équivalence en ISO 8859-6.

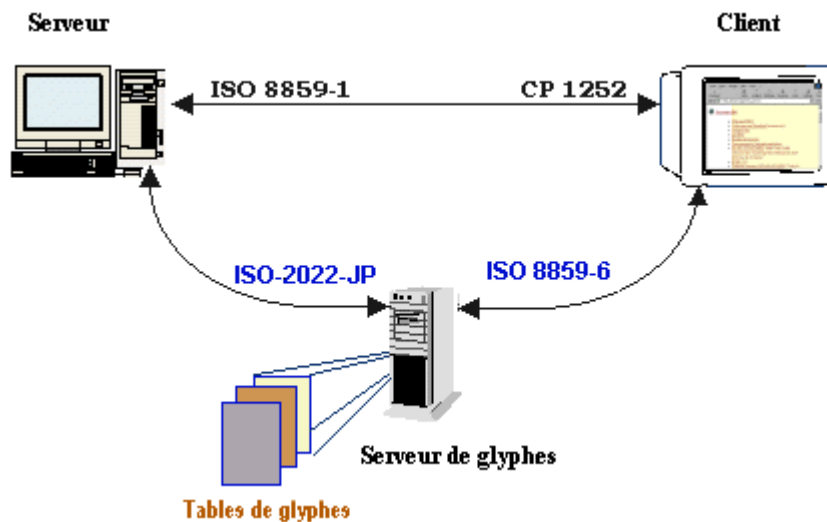


Figure 69 : Modèle de serveurs de glyphes intermédiaires

Bien que ce procédé soit lourd à gérer étant donné l’investissement énorme en terme de ressources programmables ainsi qu’en terme de ressources matérielles et logicielles impliquées, il a l’avantage de résoudre des problèmes d’accès à des documents multilingues par un procédé relativement rudimentaire. En effet, quand le serveur de glyphes récupère un document dans une langue autre que celle de la requête, il parcourt le texte HTML avant de le transmettre au client et remplace tout caractère que le client ne serait pas capable de déchiffrer par une balise qui pointerait vers son équivalent graphique sur le serveur de glyphe. Le client recevrait le texte HTML modifié à travers lequel il récupérerait les images des caractères pour en faire une composition de texte en mode graphique. Cette expérience, malgré ses inconvénients relatifs à la reconnaissance du texte pour des besoins d’édition, de recherche et de traitement électronique des données, est actuellement menée par la bibliothèque CIILIB (Character to Inline Image Conversion Library) sous forme d’une conversion des caractères japonais, cyrilliques, hébreux, chinois et coréens en vignettes graphiques à partir de documents HTML en ligne sur le serveur proxy « [DeleGate](http://193.68.137.61/ciisup.htm) »^[125].

À notre sens, la combinaison entre les deux modes serait également possible pour résoudre des situations où les deux techniques seraient complémentaires afin de combler les lacunes d’un transcodage partiel par des équivalents de caractères en mode graphique.

[125] voir URL : <http://193.68.137.61/ciisup.htm>

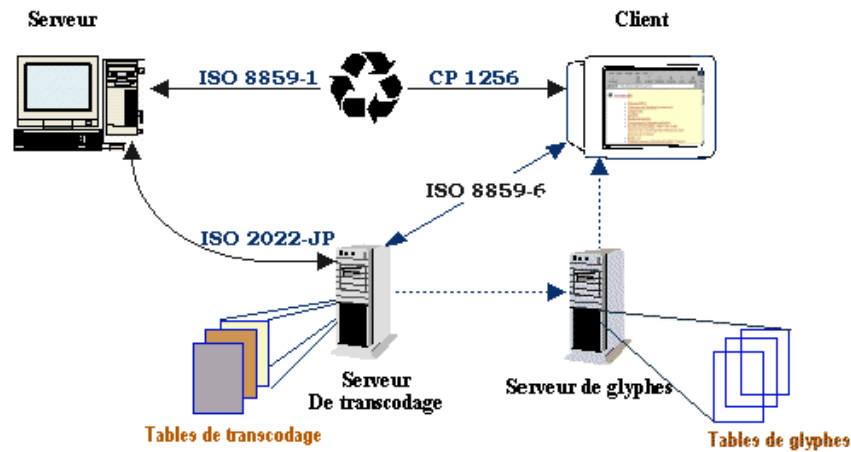


Figure 70 : Modèle combiné de serveurs intermédiaires de transcodage et de glyphes

Dans l'exemple suivant, une opération de transcodage directe et implicite a eu lieu entre le client qui affiche la ressource en Latin-1 et le serveur qui la diffuse en Windows CP 1252. Par contre, le transcodage des éléments de codes Windows CP 1256 (pour l'arabe) vers le Latin-1 a engendré, en l'absence d'une fonte et d'une police arabe, la chaîne de caractères illisibles.

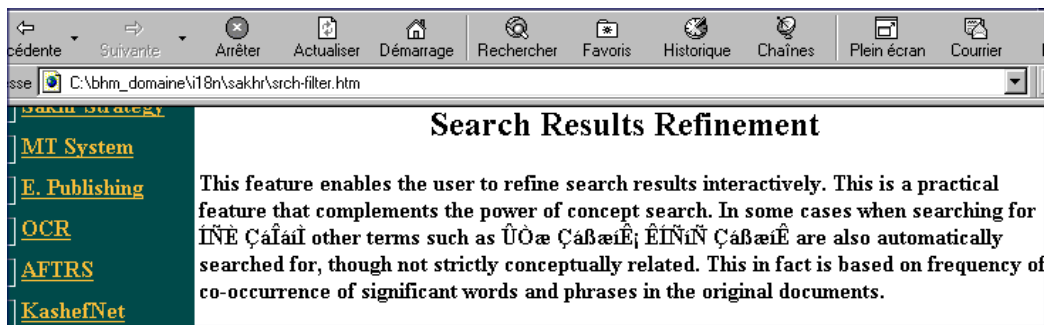


Figure 71 : Le rendu d'une conversion partielle entre jeux des caractères codés

La redirection du document sur un serveur de glyphes aurait fourni des équivalents graphiques aux caractères codés en CP 1256 pour donner un rendu identique à celui affiché dans l'exemple suivant. Il est évident que ce cas de figure en particulier pourrait être résolu selon l'une des deux méthodes suivantes :

1. Disposer d'une plate-forme de travail bilingue arabe/latin conforme à celle du serveur ;
2. Installer une police de caractères arabes et un mécanisme de gestion BIDI au niveau du client.

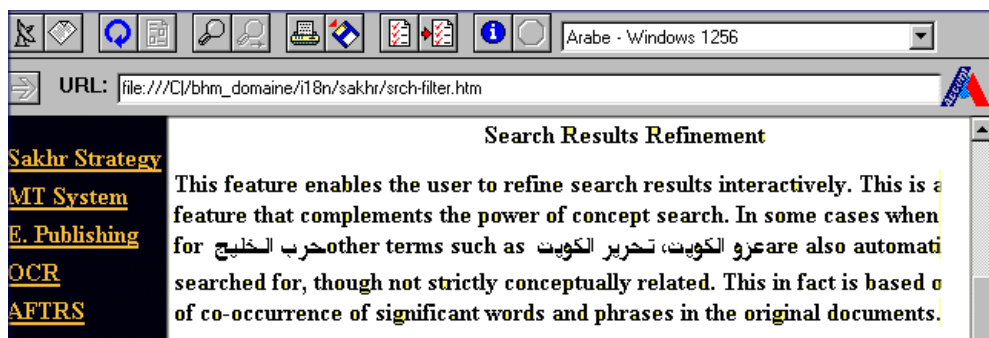


Figure 72 : Transcodage intégral entre client et serveur de profil linguistique identique

4.2.2.3.2 Les techniques de l'alignement et de l'hypertexte parallèle

Les deux situations évoquées dans les deux exemples précédents, constituent toutefois des solutions intermédiaires de bas niveau dont l'utilité occasionnelle est même dépassée par les enrichissements que ne cessent d'accomplir les efforts de l'i18n.

La complexité du multilinguisme des ressources Internet ne se limite plus uniquement à la résolution des valeurs de forme de données ou de la technique de l'encodage des caractères au moment de leur transmission entre clients et serveurs. Le gain de performance en matière de ressources et d'outils de traitement et de gestion des données électroniques, permet désormais d'approcher le concept de l'i18n par de nouvelles méthodes de travail. L'ingénierie linguistique et les techniques de la « traductique » pénètrent progressivement l'univers informationnel pour permettre de contourner les contraintes du codage des caractères par les solutions de l'analyse linguistique et de la traduction « à la carte » (traduction humaine, traduction assistée par ordinateur et traduction machine). Le balisage HTML s'en trouve impérativement amélioré pour permettre des nouveaux modes de transaction et d'échange par documents ou blocs de texte parallèles et alignés. La technique de la traduction et des liaisons réversibles se trouvent ainsi réunies pour permettre l'accès et l'échange des documents multilingues homogènes ou composites. Qu'en est-il au fait de l'échange ou de l'accès à des documents en langue arabe ?

Nous aborderons cette question en faisant référence à une approche de traitement et d'échange de documents multilingues dans le cadre d'un processus connu sous l'acronyme "WInter" ou l'i18n du Web (Web Internationalization) qui s'accomplit par un mécanisme de traduction (prédéfini ou à la volée) décrit par Tomas Carrasco Benitez^[126] comme l'une des alternatives pour résoudre les deux problématiques de la traduction et de la translittération^[127]. Dans ces deux cas de figure, le système est appelé à gérer un processus d'invocation de l'acte de traduire (ou de translittérer), de veiller à son exécution et de procéder ensuite à la mise en place du mécanisme de parallélisme et d'alignement textuel entre les deux versions linguistiques d'un même document^[128].

a. Parallélisme et alignement textuels

Par textes parallèles il est fait référence à deux ou plus versions linguistiques d'un même document textuel. La charte des Nations Unies en arabe et en anglais sont deux textes parallèles d'un même document que l'utilisateur pourrait visualiser dans deux fenêtres différentes dans lesquelles le texte défilerait dans un parallélisme synchronisé. En effet, la synchronisation entre deux textes parallèles se fait observer au moment de l'accès ou de la navigation dans une version ou dans une autre du document traduit. Une synchronisation (alignement) au niveau des points de liaison dans les deux versions linguistiques est alors observée comme critère de qualité dans la structuration des textes parallèles.

La qualité de l'alignement des textes parallèles est évaluée en fonction de la profondeur de l'équivalence textuelle. Les textes parallèles sont généralement structurés sur 4 niveaux d'alignements ([fig.73](#)) :

[126] Manuel Tomas CARRASCO BENITEZ. WInter (Web Internationalization & Multilinguism). WInter Group, 1996. <http://www.crphl.lu/~carrasco/winter/draft0.html>

[127] Web Internationalization & Multilingualism Symposium. Seville, 20-22 November 1996. <http://www2.echo.lu/oii/en/w3c-int.html>

[128] Manuel Tomas CARRASCO BENITEZ. On the multilingual normalization of the Web. Third International World-Wide Web Conference. 10-14 Avril 1995. Darmstadt, Germany. <http://www.sil.org/sgml/carrasco.html>

1. **Niveau du document** : c'est le cas le plus simple étant donné que l'identification du rapport parallèle entre deux documents pourrait se faire à un niveau externe par un mécanisme de relation entre fichiers ;
2. **Niveau de paragraphes** : ceci constitue un niveau plus avancé de complexité mais qui reste toujours réalisable grâce à la structure interne des documents qui identifie les blocs de paragraphes par les fonctions CR et LF (retour à la ligne et lignes vides) qui les séparent ;
3. **Niveau des phrases** : c'est le mécanisme le plus recommandé dans la juxtaposition des textes parallèles. Il répond aux mêmes critères que le niveau précédent partant du fait que la structure de la phrase dans un texte est définie par la présence des délimiteurs de ponctuation comme la virgule, le point-virgule et le point. Bien que ces délimiteurs ne disposent pas d'une grande valeur syntaxique, ils permettent d'identifier des objets linguistiques ;
4. **Niveau terminologique** : c'est le cas le plus complexe car la structuration du texte par séparateurs blancs (espacements) entre les mots est très vague donc démunie de toute valeur linguistique. La technique du balisage est donc nécessaire à ce niveau pour garantir un parallélisme (alignement) cohérent entre les textes parallèles.

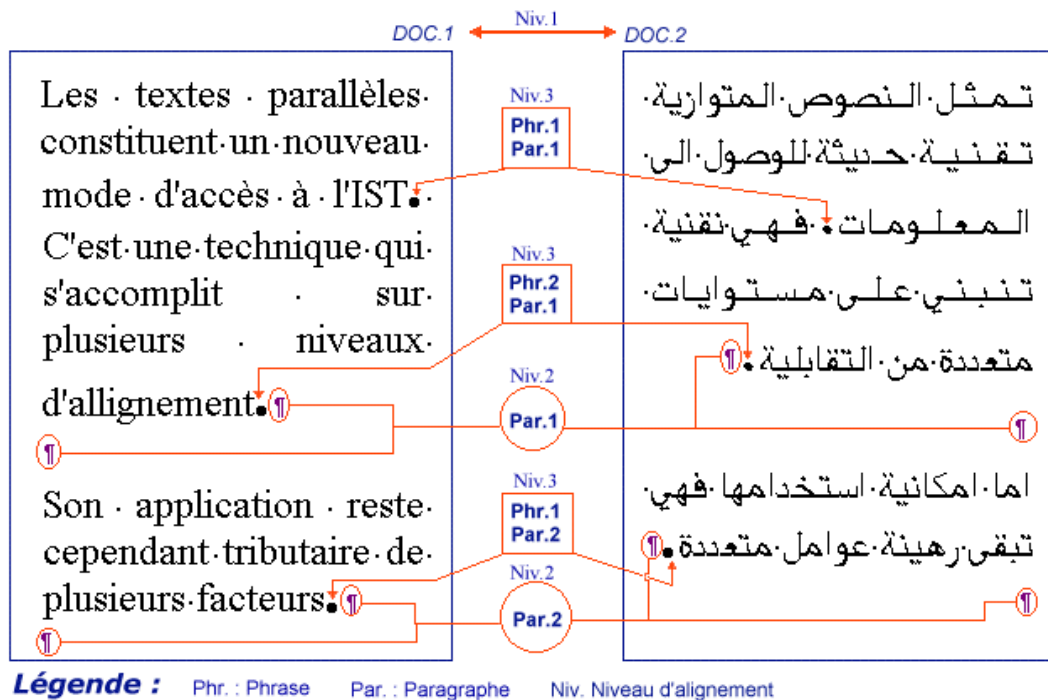


Figure 73 : Les trois premiers niveaux d'alignement des textes parallèles

Un exemple représentatif (bien que dans une autre optique de navigation) de cette technique de parallélisme ou d'alignement parallèle entre blocs de textes est visible sous l'interface Microsoft Word (en mode affichage normal) au niveau de la zone du corps du texte qui défile en mode synchrone avec le déplacement dans la zone des notes de bas de pages.

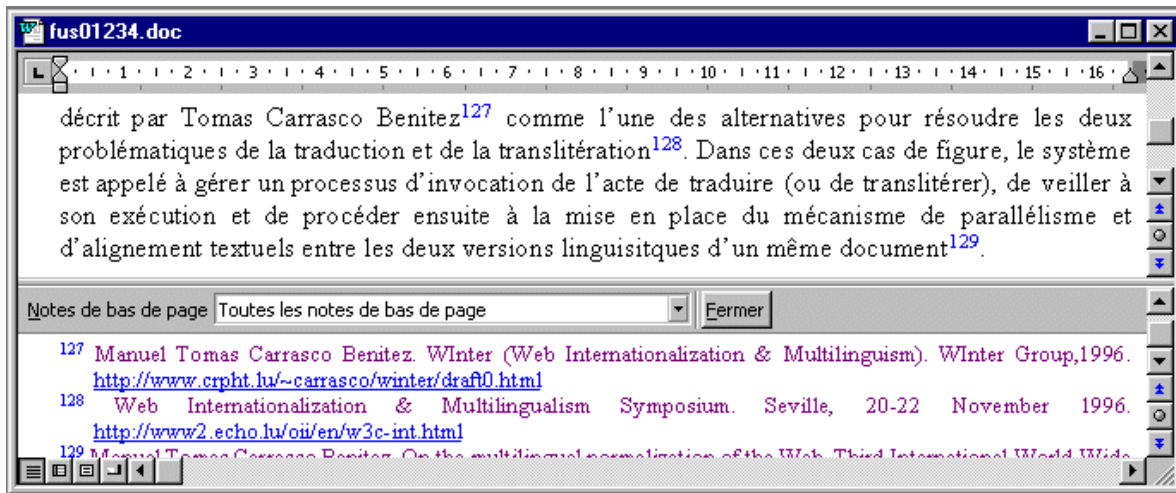


Figure 74 : Alignement parallèle entre notes et appels de notes sous l'interface Microsoft Word

Bien que cette technique constitue une étape d'évolution importante dans le traitement et l'accès aux documents multilingues, son application sur des systèmes locaux et centralisés ne constitue pas une complexité majeure ni au niveau de la structuration interne des documents pour définir les points de la symétrie textuelle ni au niveau de l'accès et de la navigation sur la base d'un affichage parallèle. Cette flexibilité relative est issue du fait que la symétrie entre les documents parallèles sur une plate-forme homogène est généralement statique et conforme à des ressources de traitement technique et linguistique prédéfinies sur le système centralisé sur lequel elles tournent.

Le parallélisme des données textuelles devient plus complexe quand il est soumis à deux facteurs nouveaux :

- ❑ un éclatement des limites géographiques (systèmes répartis et distribués) qui implique le recours aux techniques de l'hypertexte, des GUIs^[129] et des liens dynamiques sur les bases de l'architecture client-serveur ;
- ❑ une dynamique exogène de paramétrage (choix de parallélisme entre versions linguistiques exprimé par l'utilisateur au moment de l'accès). Ceci engendre la génération instantanée des textes parallèles par un mécanisme de traductique à la carte. Il n'est certainement pas possible (du moins dans l'état actuel des choses) que tout document accessible en ligne dispose de toutes les versions linguistiques imaginables. Des solutions de traductions assistées par ordinateurs sont progressivement proposées par l'industrie de la langue et l'ingénierie linguistique.

b. L'hypertexte parallèle distribué

Les concepts de l'hypertexte parallèle et de l'alignement textuel sont en réalité deux formes complémentaires du mécanisme de la liaison par balisage HTML exploitées dans une perspective de transparence d'accès et de diffusion des ressources multilingues.

L'hypertexte parallèle, comme le souligne CARRASCO^[130], est une extension du paradigme de la technique de l'hypertexte vers les langues naturelles. Selon cette définition, un utilisateur qui cherche un document en anglais pourrait recevoir en réponse un document en arabe de façon

[129] Graphic User Interfaces

[130] Manuel Tomas CARRASCO BENITEZ. WInter (Web Internationalization & Multilingualism). Op.Cit

transparente à travers un menu linguistique déroulant au lieu d'un lien hypertexte incorporé dans la version anglaise du même document. Dans ce cas de figure, le Web est censé identifier les versions linguistiques d'un même document et par conséquent établir des liens systématiques vers les uns et les autres de façon automatique ou paramétrée. Les mêmes propriétés de l'alignement des textes parallèles citées auparavant s'appliqueraient à l'hypertexte parallèle.

La dynamique des liens en hypertexte, comme définie par le langage HTML, est basée sur le principe de la liaison entre une "Ancre" et une "Cible" (Anchor & Target). La liaison sous HTML peut être accomplie par une variété de techniques (quatre) et de paramètres qui identifient l'ancre et la cible et gèrent la liaison entre elles. À part les attributs <SRC> et <ISMAP> utilisés pour des liens sur l'élément image , les éléments <A> et <LINK> en constituent une paire stratégique pour mettre en place le mécanisme de la liaison dynamique. À ces deux éléments sont ajoutés les attributs "REL" et "REV" qui spécifient davantage la nature des liens établis par les deux éléments précités. Malgré la restriction d'usage de ces deux attributs et leur acceptabilité encore contestée, certains navigateurs, comme celui développé par la Santa Cruz Operation, Inc. (SCO) sur le modèle Mosaïc, mettent en pratique ces attributs conjointement avec l'élément <LINK>.

L'élément <LINK>

Une ancre peut être spécifiée par l'élément <LINK> conformément à l'alinéa 12.3 de la norme HTML 4.0^[131] en vue d'établir un lien hypertexte entre un document HTML et toutes autres ressources ou objets identifiés. L'élément <LINK> est considéré comme un élément d'information META donc incorporé en nombre illimité dans l'en-tête d'un document HTML pour indiquer à l'agent utilisateur la présence d'un lien supplémentaire du document affiché vers d'autres ressources parallèles. Seulement, les navigateurs actuels se limitent à afficher les valeurs incluses dans le corps du document sans aucune indication sur les contenus des en-têtes qui restent exclusifs aux opérations de traitement interne (négociation linguistique, indexation par les robots, etc ...). L'une des formes d'usage de l'élément <LINK> peut être représentée comme suit :

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
    "http://www.w3.org/TR/REC-html40/strict.dtd">
<HTML>
<HEAD>
  <TITLE>Chapitre 2</TITLE>
  <LINK rel="Index" href="index.htm">
  <LINK rel="Next" href="Chapter3.htm">
  <LINK rel="Prev" href="Chapter1.htm">
</HEAD>
...le reste du document ...
```

L'élément <LINK> est enrichi, selon HTML 4.0 par des attributs qui jouent des rôles complémentaires. Nous en retenons les trois suivants : "Rel", "Rev" et "Charset".

Les attributs "Rel" et "Rev", associés à l'élément <LINK>, spécifient la nature du rapport entre la ressource cible et le document courant. L'attribut "Rel" annonce un lien suivant alors que l'attribut "Rev" définit un lien réversible en amont du document courant.

L'attribut "Charset" est également stratégique dans le contexte des documents multilingues. Puisque les liens dans un contexte ouvert et distribué peuvent pointer vers des ressources codées

[131] About the HTML 4.0 specification. W3C Working Draft 17-Sep-1997. <http://www.w3.org/TR/WD-html40-970917/>

dans des jeux de caractères différents, l'élément <LINK> (ainsi que l'élément <A> décrit ci-après) supportent l'attribut "Charset" qui permet aux auteurs des ressources de renseigner l'agent utilisateur sur la nature de l'encodage des données dans les documents liés. Ces valeurs sont également importantes pour la localisation des ressources par les moteurs de recherche.

L'exemple suivant^[132] illustre la valeur que les indications linguistiques ainsi que les attributs des liens peuvent avoir dans l'identification des ressources par les robots et les moteurs de recherche. Dans cet exemple, l'attribut <HREFLANG>^[133] est utilisé pour indiquer au moteur de recherche l'emplacement des versions allemande, portugaise, arabe et française d'une ressource quelconque.

Il est à noter que dans cet exemple, l'usage des attributs <DIR> et <CHARSET> au niveau de la section arabe indiquerait à l'agent utilisateur à la fois l'orientation du texte (RTL=Right To Left) et le jeu de caractères codés utilisé dans le document référencé (ISO 8859-6 = ASMO 708). L'usage de l'attribut <LANG> dans la section française indique que la valeur de l'attribut <Title> de l'élément <LINK> est en français contrairement aux autres sections qui prennent par défaut la valeur "en" de la langue anglaise.

```
<HEAD>

<TITLE>The manual in English</TITLE>

<LINK title="The manual in Dutch"
  type="text/html"
  rel="alternate"
  hreflang="nl"
  href="http://someplace.com/manual/dutch.html">

<LINK title="The manual in Portuguese"
  type="text/html"
  rel="alternate"
  hreflang="pt"
  href="http://someplace.com/manual/portuguese.html">

<LINK title="The manual in Arabic"
  dir="rtl"
  type="text/html"
  rel="alternate"
  charset="ISO-8859-6"
  hreflang="ar"
  href="http://someplace.com/manual/arabic.html">

<LINK lang="fr" title="La documentation en Fran&ccedil;ais"
  type="text/html"
  rel="alternate"
  hreflang="fr"
  href="http://someplace.com/manual/french.html">

</HEAD>
```

Étant donnée l'importance de ces indications supplémentaires à l'utilisateur, et en fonction de l'enrichissement continu de HTML par les attributs de l'i18n, HTML 4.0 recommande que les agents utilisateurs agissent intelligemment pour éviter l'affichage des caractères illisibles par

[132] Empruntés de : About the HTML 4.0 specification. W3C Working Draft 17-Sep-1997. Op.cit

[133] L'attribut <hreflang> fournit à l'agent utilisateur une indication sur la langue de la ressource ciblée par un lien hypertexte. L'attribut <LANG> en fait de même avec les éléments et les attributs internes au document courant.

l'incorporation d'une technique de reproduction des valeurs des métadonnées de l'élément <LINK> sous forme de :

- ❑ une icône sur la barre d'outils de l'interface utilisateur qui afficherait par simple clic en format tableau l'ensemble des ressources reliées et détectées par l'agent utilisateur ;
- ❑ un menu déroulant dans la ligne de commandes du navigateur qui afficherait en mode listing l'ensemble des ressources reliées et détectées ;
- ❑ une indication du niveau d'alignement (paragraphe, phrase, mot, etc)
- ❑ une fenêtre supplémentaire avec les URIs voire même les contenus des documents des références liées qui défilent en mode synchrone (proposition de Murray Maloney^[134]) ;
- ❑ un message d'erreur, en cas de la non disponibilité des ressources pour un affichage adéquat, destiné à informer l'utilisateur que le document existe sans pourtant pouvoir l'afficher.

L'élément <A>

L'élément <A> est utilisé pour indiquer le début d'une ancre ou la terminaison d'une cible dans un lien hypertexte dans le corps du document HTML. Le lien hypertexte décrit par l'élément <A> est généralement démarqué par une forme visuelle distinctive (mot ou phrase soulignés, image encadrée...) dans la zone de texte de l'agent utilisateur.

Un élément <A> est généralement utilisé pour définir une ancre ou une cible d'un lien hypertexte. Les deux modes d'usage les plus courants qui correspondent à ces deux aspects sont comme suit :

Accompagné de l'attribut <HREF> l'élément <A> indique un lien hypertexte vers une cible interne ou externe :

```
<A HREF="target">Objet interne du lien</A>
```

```
<A HREF="http://www.domaine.nom/document1.html#ancre1">Objet externe  
du lien</A>
```

Accompagné de l'attribut <NAME>, l'élément <A> indique une ancre qui constituerait une cible d'un lien hypertexte externe :

```
<A NAME="ancre1">Ancre1 dans le document courant</A>
```

L'élément <A> partage les mêmes valeurs linguistiques que l'élément <LINK> et dispose des mêmes attributs. Leur usage est souvent similaire et complémentaire dans la construction des liens linguistiques.

c. La négociation linguistique par versions linguistiques parallèles

Les éléments <A> et <LINK> permettent en effet de bien contrôler la liaison entre un document quelconque et ses versions linguistiques multiples. L'identification et l'accès à des différentes versions d'un document courant constituent l'objectif essentiel de la négociation linguistique par les éléments <A> et <LINK>.

[134] Murray MALONEY. Hypertext links in HTML. 1995. <http://www.sil.org/sgml/maloneyRelationships.html>

Tout document qui inclurait à son début la forme suivante des liens par l'élément <LINK> rendrait possible l'identification et l'accès à ses propres versions linguistiques par l'intermédiaire des agents utilisateurs qui auraient déjà apporté les modifications nécessaires à leurs interfaces utilisateur conformément aux caractéristiques du traitement des métadonnées exposées auparavant :

```
<BASE "http://www.myco.org/pub/subject/en/myfile.htm">
<LINK name=author href="mailto:author@myco.com" title="Author" rev=made>
<LINK name=spanish href="../sp/fichier.htm" title="Español" rel=translation>
<LINK name=arabic href="../ar/fichier.htm" title="Arabic" rel=translation>
<LINK name=french href="../fr/fichier.htm" title="Français" rel=translation>
<LINK name=german href="../de/fichier.htm" title="Deutsch" rel=translation>
```

Dans le cas contraire, et jusqu'à la mise à niveau des navigateurs pour l'affichage des éléments métadonnées sous forme d'icônes ou menus déroulants, la seule alternative est de convertir ces liens sous forme d'ancres à incorporer au sein même du corps du document courant selon la structure suivante :

```
<A name=spanish href="../sp/fichier.htm" title="Espanol" rel=translation>
[Español]</a>
<A name=arabic href="../ar/fichier.htm" title="Arabic" rel=translation>
[عربية]</a>
<A name=french href="../fr/fichier.htm" title="Francais" rel=translation>
[Français]</a>
<A name=german href="../de/fichier.htm" title="Deutsch" rel=translation>
[Deutsch]</a>
```

Ceci donnerait le rendu suivant :



Figure 75 : Le mécanisme de liens par ancrage multilingue

Une fois que les nouveaux paramètres sont acceptés par les navigateurs HTML, l'exemple précédent pourrait être enrichi comme suit et repositionné au niveau de l'en-tête :

```
<A name=spanish href="../sp/fichier.htm" title="Espanol" rel=translation
lang=sp class=automatic>
[Español]</a>
```

```
<A name=arabic href="../../../ar/fichier.htm" title="Arabic" rel=translation
lang=ar class=automatic>
[عربية]</a>
```

```
<A name=french href="../../../fr/fichier.htm" title="Francais" rel=translation
lang=fr class>manual>
[Français]</a>
```

```
<A name=german href="../../../de/fichier.htm" title="Deutsch" rel=translation
lang=de class=semi-automatic>
[Deutsch]</a>
```

L'élément <CLASS> est utilisé dans ce cas concret pour indiquer si la traduction du document original s'est accomplie automatiquement par un programme, manuellement par un traducteur ou de manière semi-automatique par une intervention humaine sur une traduction automatique.

Le contexte de la négociation par versions linguistiques parallèles obéît en fait à un enchaînement de procédures et un ensemble de facteurs dont dépendra la bonne mise en marche du mécanisme en question. Nous en citons : la génération des versions linguistiques parallèles ; la structuration des versions linguistiques parallèles ; la stratégie de la liaison entre deux ou plus versions linguistiques parallèles ; la requête et la négociation des liens linguistiques parallèles par interface utilisateur ; la traductique et la génération automatique des textes traduits.

La génération des textes parallèles :

La génération des versions linguistiques des documents sur Internet (en simultanée ou en différé) peut se faire selon trois principes complémentaires :

- ❑ Une traduction machine intégrale : les seuls cas qui enregistrent un taux élevé de pertinence sont ceux qui intègrent la technique de l'Intelligence Artificielle (IA) ou ceux qui partent d'un corpus linguistique dans des bases de données spécialisées. Les choix des équivalences linguistiques sont prédéfinis et l'ambiguïté sémantique et la redondance morphologique sont traitées et éliminées à l'avance. C'est donc l'utilisateur qui déterminera l'efficacité du système par la nature des données qu'il soumet à la traduction automatisée intégrale.
- ❑ Une traduction semi-automatique : C'est la catégorie la plus répandue et connue sous l'appellation "Traduction Assistée par Ordinateur". L'outil de traduction est soumis à un paramétrage aléatoire qui nécessite l'intervention successive de l'utilisateur pour établir un choix entre deux termes ou plus. À l'opposé des systèmes IA, les outils de la TAO sont généralement démunis de toute performance d'analyse sémantique et se contentent dans les meilleurs des cas à gérer une concordance syntaxique de bas niveau.
- ❑ La traduction humaine : avant d'arriver à l'ère de la machine biologique (certaines réalisations laisseraient comprendre que nous y sommes déjà), l'équilibre Homme-Machine n'est pas encore atteint, tout du moins sur le plan cognitif. Il n'y aurait pas mieux que l'Homme qui comprendrait l'Homme « à sa juste humeur » dirions nous. La traduction humaine resterait le choix des spécialistes dans des domaines de pointe ou dans des langues à flexion énorme comme le géorgien ou l'arabe.

La structure hypertexte parallèle

Comme tout document HTML, une structure hypertexte parallèle d'un document hypertexte parallèle est composée de deux parties, un en-tête et un corps de document. L'en-tête contient les

éléments descriptifs du document (les métadonnées). Le corps inclut les données elles-mêmes organisées en sections monolingues selon la variation linguistique du document concerné.

L'en-tête est généralement un fichier HTML qui remplit les deux fonctions suivantes :

- Permettre à l'utilisateur de choisir une version linguistique ;
- Être utilisé convenablement par les applications Web multilingues (Winter) ;
- Indiquer la localisation physique de toutes les versions linguistiques du document.

La figure suivante illustre la structure standard d'un service de diffusion des versions linguistiques d'un seul document.

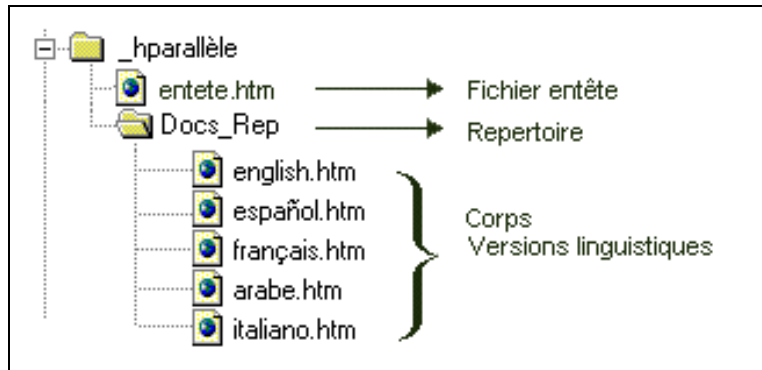


Figure 76 : Modèle standard de structuration d'un service hypertexte parallèle

La figure suivante représente par contre la structure d'un service hypertexte parallèle d'une série de documents.

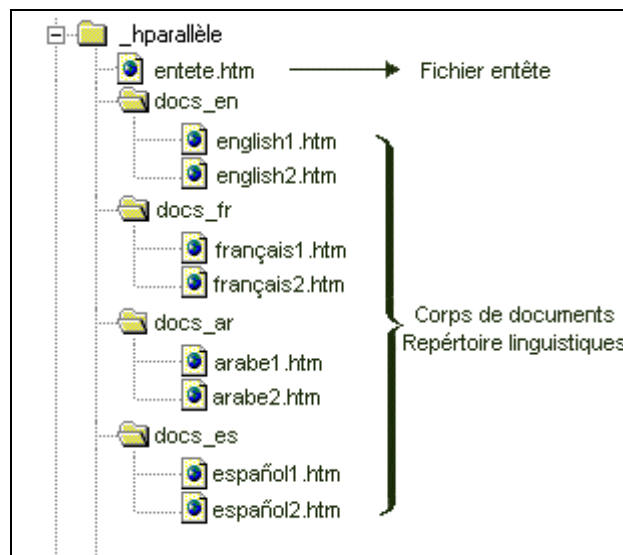


Figure 77 : Modèle de structuration d'un service hypertexte parallèle multidocuments

Le fichier « entete.htm » est le point de chute des navigateurs dans une phase d'accès aléatoire. Le même fichier constituerait la source d'une génération à la volée des versions linguistiques dans une opération de traduction instantanée. Les navigateurs multilingues devraient normalement utiliser le fichier d'en-tête de manière transparente pour accéder aux diverses versions linguistiques.

La stratégie des liaisons entre documents hypertextes parallèles

Le mécanisme de la liaison entre plusieurs versions linguistiques parallèles est basé sur le principe fondamental de la synchronisation. Chaque page dans chaque version linguistique doit avoir le même contenu et la même structure interne (texte, nombre de paragraphes etc.). La synchronisation entre les pages et les documents parallèles devrait se faire de manière automatique. L'intervention humaine ne devrait avoir lieu que pour affiner l'édition ou la présentation du document obtenu.

La stratégie de la liaison entre versions linguistiques devrait également viser un minimum d'efforts de maintenance. La solution que propose Carrasco dans ce sens est de faire usage des deux modes de liens explicites et implicites. Les liens explicites sont physiquement présents dans le documents à l'instar de ceux dont il est fait usage dans les documents HTML courants (balises du genre "<A...>Textes"). Les liens implicites sont par contre calculés par des mécanismes d'analogie entre les degrés d'alignement des textes.

Dans le cas d'un parallélisme étendu entre des versions linguistiques multiples des documents volumineux, il est recommandé de définir une seule version avec des liens explicites. L'usage de toutes les autres versions linguistiques avec des liens implicites s'effectuerait sur le seul référent du document principal.

Seulement, le degré d'alignement devrait être calculé à l'avance et enregistré dans la structure des données hypertextes parallèles. La détermination des niveaux d'alignement serait tributaire du plus bas niveau d'alignement. Deux textes alignés au niveau des paragraphes ne permettraient pas de calculer les liens implicites au niveau des phrases. Deux textes alignés au niveau hiérarchique supérieur des documents ne pourraient avoir de liens implicites qu'au niveau du début et de la fin des documents.

Il est admis dans ce sens qu'un alignement simple ne pourrait pas descendre vers un niveau plus bas que la structure de la phrase car les mots qui la constituent ne disposent pas d'un balisage à valeurs syntaxique ou sémantique quelconques (fig.73). Les espaces qui séparent les mots n'ont pas cette valeur dont jouissent les signes de ponctuation, les retours à la ligne et l'insertion des lignes vides (RC/LF).

Requête et négociation des versions linguistiques des documents en hypertexte parallèle

La requête et la négociation des versions linguistiques des documents parallèles relèvent tout directement du domaine des Interfaces Homme-Machine.

L'interface cliente devait être capable de revendiquer une version linguistique d'un document dans au moins les deux formes suivantes :

- ❑ Le document est choisi en fonction d'un ordre linguistique présenté sous forme d'une liste de préférences proposées sur l'interface de l'agent utilisateur. Le serveur répondra en soumettant la version la plus proche de la tête de liste des langues formulées par l'utilisateur, ou le cas échéant, celle qui se rapproche le plus de l'ordre préférentiel par défaut dont il dispose. Une indication des autres versions disponibles est également soumise à l'utilisateur pour information ;

- ❑ Un document est sollicité dans une version linguistique bien particulière. Le serveur est censé transmettre uniquement la version demandée par l'utilisateur accompagnée d'une liste des versions disponibles.

Dans un cas comme dans un autre, le recours aux champs « Accept-Language » et « Content-Language » (fig.67) renforcerait davantage le mécanisme des hypertextes parallèles. Un champ « Accept-Language » avec une seule variable linguistique imposerait une exclusivité linguistique au niveau de la réponse. Un champ « Content-Language » de la part du serveur renseignerait le client sur les versions linguistiques disponibles. Ce champ devrait avoir une équivalence graphique dynamique sur l'interface utilisateur pour informer l'utilisateur de toutes les versions linguistiques de tout document demandé.

La traduction parallèle

Nous voyons de plus en plus fréquente l'intégration des techniques de l'ingénierie linguistique et du traitement des langues naturelles dans l'univers Internet bien qu'à des niveaux de perfectionnement encore contestés. Les outils de la traduction automatique des ressources HTML, sont encore à leurs débuts et ne peuvent en aucun cas remplacer les techniques assez lourdes et perfectionnées de l'industrie de la langue et de l'intelligence artificielle. Ailleurs où le traitement automatisé des langues fait l'objet d'une grande intensification d'outils sophistiqués pour la traduction (dictionnaires électroniques bilingues, bases de données lexicales, analyseurs et générateurs morphosyntaxiques, systèmes d'analyse textuelle, systèmes de traitement de corpus multilingues, thésaurii multilingues, navigateurs syntaxiques, "désambiguisateur" de sens ...), la traduction à la carte de documents retrouvés par un processus d'accès Internet nécessiterait la mise en action de tous ces artefacts linguistiques au même moment pour générer à la volée une équivalence linguistique dans le choix exprimé de façon aléatoire par l'utilisateur au moment de la consultation.

Seulement, une traduction automatisée, selon LABED^[135], se fait généralement sur trois étapes d'un niveau de complexité qui nécessite des opérations de prétraitement minutieuses :

- ❑ l'analyse du texte source qui aboutit à une représentation profonde en langue source dans laquelle le texte est analysé en une structure détaillée et non de surface constituée en *"unités lexico-syntaxiques représentées sous différentes formes de structures autorisées"*^[136]. Cathérine Lelou développe cette opération en précisant que l'aboutissement à cette étape d'analyse du langage devrait passer par les dictionnaires des réseaux sémantiques et des moteurs de traitement du texte qui permettrait d'accomplir les quatre phases d'analyse à savoir l'analyse morphologique, l'analyse syntaxique, l'analyse phonétique et l'analyse sémantique du langage^[137].
- ❑ le transfert ou la conversion de cette représentation profonde interne vers une langue cible. Le transfert *"consiste à identifier, à partir des descripteurs ou composants du texte dans la langue source, les descripteurs correspondants dans la langue cible. Les descripteurs de textes sont les unités lexicales et les unités syntaxiques"*^[138].

[135] Lamia LABED, Ridha SALHI, Salem GHAZALI. Génération de textes dans un système de traduction anglais-arabe : Torjmane. In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.91-114.

[136] Lamia LABED, Ridha SALHI, Salem GHAZALI. Génération de textes dans un système de traduction anglais-arabe : Torjmane. Op.Cit. p.96.

[137] Cathérine LELOU. Moteurs d'indexation et de recherche : environnement client-serveur, Internet et Intranet. Paris : Eyrolles, 1998. 277 p.

[138] Lamia LABED, Ridha SALHI, Salem GHAZALI. Génération de textes dans un système de traduction anglais-arabe : Torjmane. Op.Cit.p.96.

- la génération définitive du texte dans la langue cible à base des résultats de transfert.
« Dans la phase de génération, la séquence d'étiquetage générée par le transfert est convertie dans la langue cible ; c'est-à-dire que le texte cible est généré à partir des descripteurs obtenus dans la phase de transfert. La phase de génération est symétrique à la phase d'analyse »^[139].

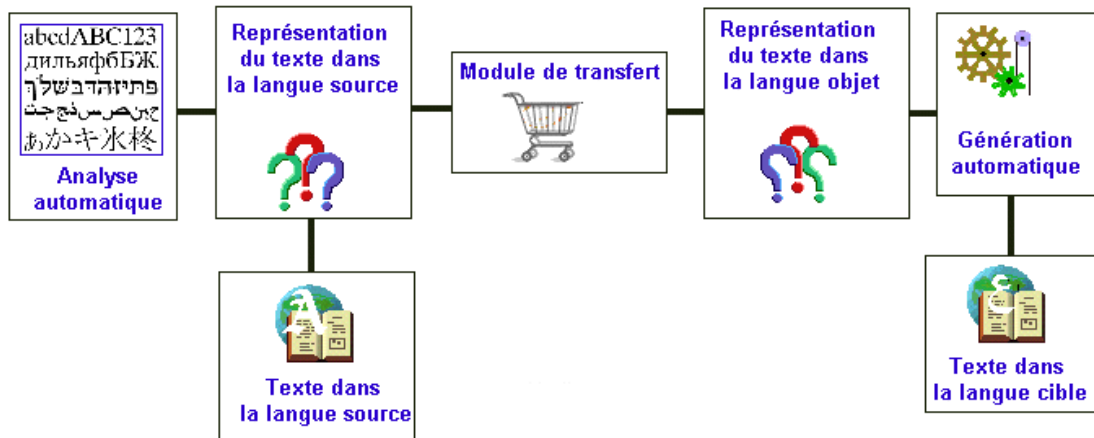


Figure 78 : Les phases du processus de traduction automatisée

Inspiré de : Lamia LABED, Ridha SALHI, Salem GHAZALI. Génération de textes dans un système de traduction anglais-arabe. Op.Cit.p.95

Les travaux sur la langue arabe dans ce contexte particulier ont fait surgir une multitude de problèmes inhérents à sa structure inflexionniste de haut niveau ainsi qu'à une série de contraintes dans la phase d'analyse linguistique du point de vue des traitements des noms composés, des pronoms, des groupes adjectivaux, de la traduction de types de verbes particuliers etc. Les études accomplies par l'Institut Régional des Sciences Informatiques et de Télécommunication de Tunis (IRSIT) constituent dans ce sens une référence importante pour l'approfondissement dans ce secteur de recherche spécialisée^[140].

Le mécanisme de la traduction automatisée des langues naturelles dans une perspective de traitement parallèle sur Internet, nécessiterait donc un effort de « désambiguïsation » énorme des portées conceptuelles des termes choisis ou de recherche face à la flexibilité du langage naturel. Tout document localisé par une opération de recherche multilingue devrait normalement être traduisible selon les préférences linguistiques de l'utilisateur. Étant donné les caractéristiques inflexionniste des langues, l'ambiguïté sémantique et la complexité structurale syntaxique, cette opération nécessiterait deux pré-requis fondamentaux à savoir une spécialisation des systèmes utilisés et la négociation interactive avec l'utilisateur pour résoudre les cas d'ambiguïté aiguë conformément au profil cognitif et socioculturel de l'utilisateur. Ces deux aspects ne sont pas encore garantis sur Internet pour les deux simples raisons de la multidisciplinarité du réseau Internet (donc impossibilité de disposer d'outils linguistiques d'une spécialisation profonde) ainsi que pour le champ réservé des systèmes de traduction aux langues dominantes généralement à caractères latins. La traduction humaine reste encore la solution incontournable pour atteindre des degrés élevés de perfection bien que Gregor ERBACH^[141], de la *German Research Center*

[139] Lamia LABED, Ridha SALHI, Salem GHAZALI. Génération de textes dans un système de traduction anglais-arabe : Torjmane. Op.Cit.96.

[140] Voir dans ce sens le produit TORJMANE de l'IRSIT pour la traduction automatisée entre l'anglais et l'arabe.

[141] Gregor ERBACH est un chercheur dans le domaine des technologies du langage. Il a fait des études en linguistiques, informatique et intelligence artificielle et a accompli des recherches dans les domaines du formalisme et des outils du développement de la grammaire. Voir sa biographie sur l'URL : <http://lt-www.dfki.de/~gor/> et sa bibliographie sur l'URL : <http://www.coli.uni-sb.de/~erbach/schriftenverzeichnis.html>

for Artificial Intelligence^[142], cité par Martin BRYAN^[143], considère que la traduction humaine est lente et trop coûteuse par comparaison à la mise en place des sites Web unilingues.

En amont de l'opération de la traduction parallèle, l'acte de recherche est non moins important car un utilisateur potentiel qui ne comprendrait pas suffisamment une langue quelconque ne pourrait pas en faire usage pour récupérer un document écrit dans cette même langue. La recherche multilingue croisée (rechercher un document dans une version linguistique déterminée par un concept exprimé dans une autre langue) nécessiterait un dépassement de la traductique mot à mot pour aboutir à un traitement des requêtes en langage naturel. D'ailleurs, la correspondance entre mots-clés de recherche ne fonctionne pas entre disciplines ou langues diverses. Pour identifier convenablement l'équivalent d'un terme dans une langue cible, il faudrait avoir une indication parallèle de son domaine d'application et procéder à une "épuration" stylistique (métaphorique et figurative) et sémantique (synonymie, homonymie, homographie etc...) du concept ou de l'expression à traduire. La recherche en langage naturel devrait donc s'effectuer en mode arbitraire avec la possibilité de lui adjoindre des opérations de filtrage et de traitement (épuration linguistique) pour en faire une équation qui correspondrait à l'esprit de requête du genre : *"trouvez-moi des documents similaires à ceci"*.

Une recherche sur documents multilingues ne saurait cependant pas se faire à base d'une exhaustivité des équivalents linguistiques du terme utilisé. La recherche devrait alors être limitée aux langues de choix de l'utilisateur qui s'exprimerait dans l'esprit de requête suivant : *"trouvez moi des documents dans les langues suivantes qui couvriraient le même sujet que ceci"*.

Un gain de qualité dans la pertinence et la précision des réponses pourrait être assuré par l'usage des « stems », bien que cette technique ne fonctionne pas toujours avec des langues à taux d'inflexion élevés comme la langue arabe. L'expansion automatique des questions de recherche pour usages des alternatives synonymiques ou la limitation à des zones de recherche améliorerait considérablement la qualité des réponses.

d. Les systèmes parallèles : un problème de glyphes autant que de langues

La conclusion tirée de cette problématique des textes et des hypertextes parallèles centralisés et décentralisés est sans doute la problématique essentielle qui nous a guidé tout au long de notre travail : le traitement de la langue arabe comme langue alphabétique (bien que les distinctions dans le traitement automatisé entre les langues alphabétiques, idéographiques et syllabiques s'estompent rapidement) ne devrait plus représenter un handicap majeur au niveau de la codification, de l'échange et du traitement à part les deux phénomènes de la représentation graphique et de la gestion de la directionnalité droite-gauche. Les sources des problèmes majeurs dans l'incorporation de la langue arabe dans les systèmes ouverts et distribués mondiaux est aujourd'hui un problème d'affichage (outre les problèmes de l'analyse linguistique dans les applications avancée). Un hypertexte parallèle entre une source latine et une cible arabe est techniquement fonctionnel sauf qu'il n'est pas symétrique. Une plate-forme arabe afficherait convenablement un texte latin relié alors que l'inverse n'est pas encore évident. La reconnaissance des caractères arabes n'est pas garantie sur toutes les plates-formes de traitement automatisé de façon implicite à la manière du caractère latin. La raison est connue et elle est exposée dans la partie introductive de ce chapitre : antécédence technologique, extension géographique et culturelle du caractère latin, son renforcement par les normes internationales...

[142] <http://cl-www.dfki.uni-sb.de/cl/registry/draft.html>

[143] Martin BRYAN. Web Internationalization & Multilingualism Symposium. Seville, 20-22 November 1996. <http://www.sil.org/sgml/seville-w3c-int.html>

La diversité des plates-formes matérielles et de leurs configurations linguistiques hétérogènes constitue également un contrepoids à la transparence linguistique universelle. Les différents cas de figure d'un échange de données multilingues sur Internet peuvent prendre forme d'une négociation client-serveur sur la base des rapports linguistiques suivants :

1. Un client unilingue ➔ un serveur unilingue ;
2. Un client unilingue ➔ un serveur multilingue ;
3. Un client multilingue ➔ un serveur unilingue ;
4. Un client multilingue ➔ un serveur multilingue.

L'un des soucis de l'i18n et du Wlintering est d'assurer que les interfaces graphiques Homme-Machine disposent au niveau des postes clients des ressources graphiques adéquates pour accomplir un rendu multilingue riche et transparent. Parmi les recommandations que nous avançons pour l'aboutissement à cet état de fait :

- ❑ développer des fontes auto-ajustables pour des situations de ligatures, de superpositions de caractères et de styles d'impression ;
- ❑ mettre en accès public des utilitaires de décodage gratuits (en mode plug-ins) ;
- ❑ mettre en place un mécanisme de téléchargement gratuit des fontes à la demande sur le Web ;
- ❑ concevoir un schéma d'adressage et d'identification de ressources des caractères ;
- ❑ constituer un dépôt public de fontes (des solutions pour les droits d'auteurs de fontes sont également à prévoir) ;
- ❑ développer et généraliser des serveurs de glyphes intermédiaires comme décrits auparavant dans ce chapitre ;
- ❑ renforcer l'usage des feuilles de style pour harmoniser la présentation des documents HTML et gérer de façon uniforme les attributs <LANG> pour l'identification des propriétés linguistiques des parties de texte des documents multilingues.

Les acquis actuels d'Internet en termes de fontes et de glyphes sont uniquement conformes aux exigences de la langue anglaise. Pour des contextes multilingues complexes, les solutions d'extension de codes et du codage sur 16 et 32 bits, comme nous l'avons exposé ci-haut, sont désormais inévitables. Un Web multilingue est supposé gérer à la fois Unicode et le jeu de caractères codés local (i.e. ISO 8859-6) avec la possibilité d'établir des correspondances entre les deux et d'exercer des ajustements au niveau des caractères non identiques entre les deux codes (i.e. un « é » en Unicode est traduit en « e » en ASCII ; un caractère non graphique est traduit en « _ »).

Le codage sur 16 bits est la source d'un paradoxe fonctionnel d'Unicode. Il en constitue le point fort car il permet de couvrir tous les caractères du monde ; il en constituent cependant le point faible car il est resté tributaire de l'ASCII et de la langue anglaise comme langue de base. Sans un mécanisme de compression, Unicode dédouble l'espace de stockage ([fig.23](#)) et le volume du temps de transmission car les premiers 8 bits ne sont pas nécessaires pour l'anglais dans une matrice 16 bits. Ils sont donc traités en tant que suites de zéros.

Le grand acquis de l'adoption d'Unicode comme jeu de caractères codés par défaut pour les documents HTML à la place d'ISO 8859-1 (latin-1), contribuerait à l'i18n intégrale des services et ressources Internet. Seulement, cette démarche resterait toujours partielle si les balises de

formatage restent limitées à l'ASCII. L'adoption d'Unicode rendrait-elle possible l'usage de tout autre jeu de caractères codés (un sous-ensemble d'Unicode) pour la représentation du texte ainsi que des littéraux de formatage ?

La réponse est d'emblée positive si l'on considère la solution déjà en usage des références numériques des caractères pour représenter les valeurs des caractères accentués ou spéciaux. À notre sens, une pareille technique appliquée au répertoire Unicode faciliterait l'identification codée en références numériques de tous les caractères du monde. L'exemple suivant illustre l'usage d'un codage des caractères ISO 8859-1 (le terme élèves) par références numériques et alternatives alphabétiques.

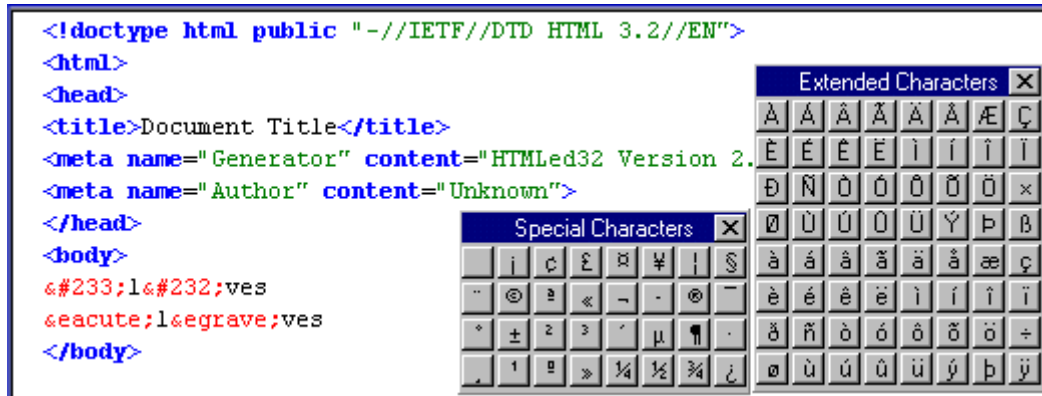


Figure 79 : Codage des caractères spéciaux et étendus par références numériques et alphabétique sous HtmlEd.

Il serait donc possible, mais par un consensus général, d'adopter cette méthode sous l'égide d'Unicode. Les références numériques resteraient non ambiguës face aux changements des jeux des caractères utilisés et aux opérations de transcodage (conversion) puisqu'elles font toujours référence aux mêmes éléments Unicode. Ceci rejoint en quelque sorte notre proposition formulée auparavant pour la conversion intégrale des URIs en codes numériques à segmentations hiérarchiques.

4.2.2.3.3 Solutions grand public Vs solutions propriétaires

Le problème des fontes et de l'affichage des caractères arabes (sans considération des réalisations progressives de l'Unicode et de l'ISO/IEC 10646) a toujours été adressé par des bouts de solutions propriétaires proposées de manières disparates et inachevées par des sociétés commerciales ou du domaine de la recherche. Le poids des grandes firmes du génie logiciel et des systèmes informatiques multilingues sur le marché mondial a toujours constitué un frein pour l'épanouissement et la mise en valeur des initiatives nationales et régionales. Le poids de Microsoft dans le monde arabe a sanctionné l'apparition d'initiatives locales pourtant assez prometteuses pour faire injecter sur le marché et dans les usages des solutions de traitement multilingues de l'information. Toutes sont vouées à s'aligner derrière le géant américain pour proposer des solutions avec un habillage Windows (§ 6.2). Le projet Alifcom^[144] de l'IRSIT^[145] pour un accès Vidéotex en langue arabe est l'un des projets contraints à migrer et rester dépendant de la version arabisée de MS-Windows.

La question est alors de savoir si l'esprit dans lequel tout Internet a été mis en place dès ses origines, à savoir la gratuité et la transparence d'accès, n'est pas en contradiction avec les modes

[144] URL : <http://www.irsit.rnrt.com>

[145] URL : <http://www.irsit.rnrt.com>

de son fonctionnement courant : une généralisation des solutions propriétaires, des mécanismes de cryptage et des droits d'accès. Les grands projets actuels de la mondialisation et de l'i18n sur la base d'une diversité linguistique (considérée enrichissante) des systèmes d'information et de communication planétaire contredisent à leurs origines l'idée de l'appropriation et de l'exclusion sur la base du monopole technologique. L'idée de développer un codage sur 16 et 32 bits est une forme de contre poids à ce penchant monopolistique.

Mais jusqu'à ce que ces solutions deviennent un soubassement uniforme de tous les systèmes en œuvre, les commerçants du génie logiciel proposent l'intermédiation lucrative pour les minorités linguistiques dans le domaine électronique. À l'instar des courtiers de l'information (brokers) dont le champ d'action et l'intérêt économique diminuent progressivement par le phénomène Internet et l'accès direct aux documents primaires, nous estimons que l'intermédiation linguistique est également vouée à s'estomper progressivement pour laisser place à l'alternative transparente. La meilleure solution serait celle d'un navigateur universel capable de s'adapter « à la volée » chaque fois qu'il rencontre des caractères d'une nouvelle écriture. C'est le principe sur lequel se basent les développeurs des ressources et des serveurs multilingues que l'on commence à voir sur les voies de l'Internet.

Nous exposons ci-après certaines des solutions de contournement à ce goulet d'étranglement linguistique qui ont été pratiquées dans des moments d'austérité technologique et qui se pratiquent encore tirant profit des innovations dans les infrastructures et les logiciels du domaine public.

a. Les caractères images

Les débuts d'Internet ont connu l'apparition des alternatives linguistiques à base d'imagerie des caractères non reconnus par les systèmes à soubassement latin. Cette tendance consiste à insérer des textes en blocs de paragraphes, de mots ou même de caractères séparés, sous forme d'images numérisées ou composées par les outils graphiques. Il est d'ailleurs encore courant de voir cette méthode utilisée pour l'insertion des grands titres d'œuvres, d'institutions ou d'individus à côté de leurs équivalents linguistiques latins en forme caractères.

Les deux inconvénients majeurs de cette méthode sont d'une part la lourdeur des images par rapport aux textes au moment de leur transmission sur les réseaux, et d'autre part l'élimination totale de toutes possibilités de traitement linguistique (indexation par les moteurs de recherche, enregistrement au format texte, édition des attributs textuels...) normalement prévue pour les données textuelles.

b. Les fontes à largeurs fixes

L'agent utilisateur doit paramétrer l'une de ses fontes (généralement la fonte à largeur fixe) sur une fonte spécifique ; celle censée être utilisée pour afficher un type particulier de caractères. Cette méthode est utilisée par des navigateurs non adaptés à l'i18n mais qui ciblent une langue particulière (généralement non latine) autre que celle par défaut installé sur leurs plates-formes locales.

À priori, l'utilisateur aurait téléchargé et installé une fonte particulière pour afficher une seule police de caractères. Cette méthode permettrait par contre de traiter une seule langue non latine à la fois. Le projet Perseus^[146] en est un exemple concret. La figure suivante illustre

[146] URL : <http://www.perseus.tufts.edu>

l'accomplissement de cette méthode sous Internet Explorer 4.0 Version latine pour afficher un texte en arabe par un paramétrage personnalisé du choix de la fonte et de langue appropriée.

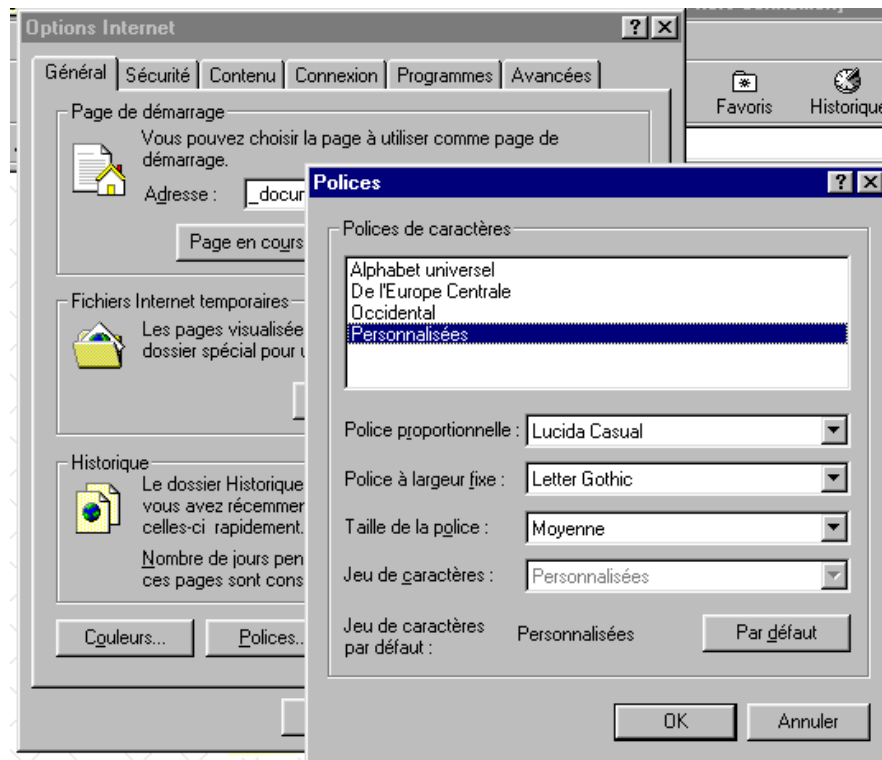


Figure 80 : Choix de fonte personnalisée sous MS-IE 40

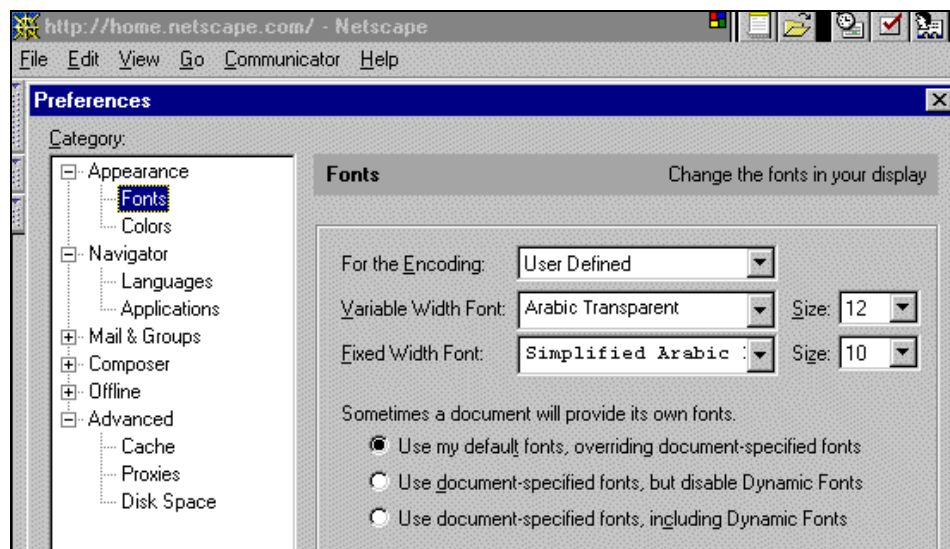


Figure 81 : Choix de fonte personnalisée sous Netscape Communicator

c. Les fontes Unicode

Les fontes Unicode peuvent être utilisées en substitue aux standards des fontes attribuées sur la base des 256 caractères codés des normes courantes à 8 bits. Avec Unicode, il n'est point nécessaire de spécifier les langues de manière explicite puisque chaque famille de langues basée sur le concept d'écriture et non de langue, dispose de ses propres points de codes. Les caractères arabes, grecques ou cyrilliques sont saisis directement par un simple changement de famille de

caractères sur l'interface utilisateur. Les exemples les plus parlants de cette liaison entre les écritures et les fontes qui leur sont associées sont les produits Unicode de la société Accent Soft. Microsoft Office Arabe en fait autant. L'exemple suivant illustre l'activation de la sélection automatique des fontes en fonction des associations qu'elles ont avec les familles d'écritures.



Figure 82 : L'association automatique entre langues et fontes dans les produits Unicode d'Accent Soft

Unicode et ISO/IEC 10646 constituent la solution futuriste par excellence. Des segments de ces deux normes peuvent toujours être téléchargés pour remplacer des entités, des tables de substitution de glyphes et/ou les glyphes eux même.

d. Le format RTF

La solution des formats des textes enrichis (rich text format) sous Microsoft Word ou WordPerfect implique l'incorporation des informations des fontes dans le fichier RTF lui même. Il serait par contre souvent nécessaire de télécharger la fonte correspondante au contenu du fichier pour la restitution du fichier dans sa forme originale. L'agent utilisateur devrait être alors configuré de sorte qu'il invoque, en tant qu'application d'aide (application helper) le programme concerné (traitement de texte) chaque fois qu'il trouve une extension « RTF » d'un fichier qu'il traite.

L'un des inconvénients majeurs du format "RTF" est l'aspect propriétaire des applications qui le gèrent malgré leur extension massive (i.e. MS Word; Word Perfect).

e. Les utilitaires de lecture des pages graphiques

Un moyen plus développé que la méthode des caractères images est l'usage des applications prévues pour la lecture des pages de texte en formats graphiques. Acrobat, Replica, Envoy^[147], ou le vieux standard Postscript, constituent aujourd'hui une forme complémentaire de diffusion des textes multilingues au delà des barrières linguistiques. S'ils répondent parfaitement à des situations d'information grand public sans besoin d'analyses textuelles complexes, ils connaissent encore des difficultés inhérentes à la lourdeur de chargement des pages ainsi qu'à l'absence d'outils de traitement textuel (recherche, conversion et édition).

[147] URL : <http://www.twcorp.com/envoy.htm>



Figure 83 : Le traitement des pages graphiques par Acrobat Reader

Les utilitaires les plus connus sont utilisés sur des domaines à accès public Acrobat Reader, Common Ground (Digital Paper), Envoy, Ghostscript et GSview (Visionneuse Postscript pour Windows), et Replica.

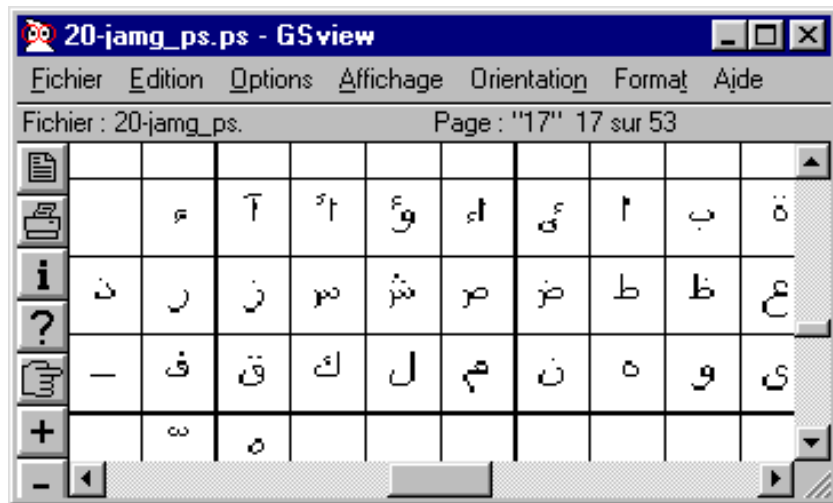


Figure 84 : La visionneuse de Postscript Gsview

f. L'attribut "FACE" de l'élément

Netscape navigator et Microsoft Internet Explorer ont ajouté un nouvel attribut "FACE" à l'élément pour permettre aux développeurs de pages HTML de spécifier leur préférence de fontes durant l'affichage d'une séquence de texte. Cette méthode, bien que déclarée par HTML 4.0 comme technique abandonnée (*deprecated*), est encore en vigueur par soucis de rétroactivité. Nous procédons à travers la série de copies d'écrans suivante à démontrer l'effet de l'attribut "FACE" dans la reconnaissance du caractère arabe dans des environnements linguistiques arabe et latin.

Test 1 : Caractères arabes dans un éditeur HTML latin et agent utilisateur latin sous Windows Arabe SANS choix explicite de fontes

Phase d'édition :

Le texte arabe est saisi dans sa forme explicite en glyphes appropriés (grâce à la plate-forme Windows arabe) ou à travers un codage par références numériques et alphabétiques. La première forme est gérée par la présence de la plate-forme arabe de Windows ; la seconde est conforme au codage interne de l'éditeur latin HTMLEdit32, raison pour laquelle les grilles des caractères étendus et des caractères spéciaux (affectées dans ce cas concret par le soubassement arabe de Windows) effectuent un codage équivalent aux caractères étendus et spéciaux latins à l'intérieur de l'application latine.

Le mode de saisie dans l'éditeur latin est par contre forcé de gauche à droite (mode logique latin : voir § [IV-2.3.1](#)) conformément à la structure interne de l'éditeur en question

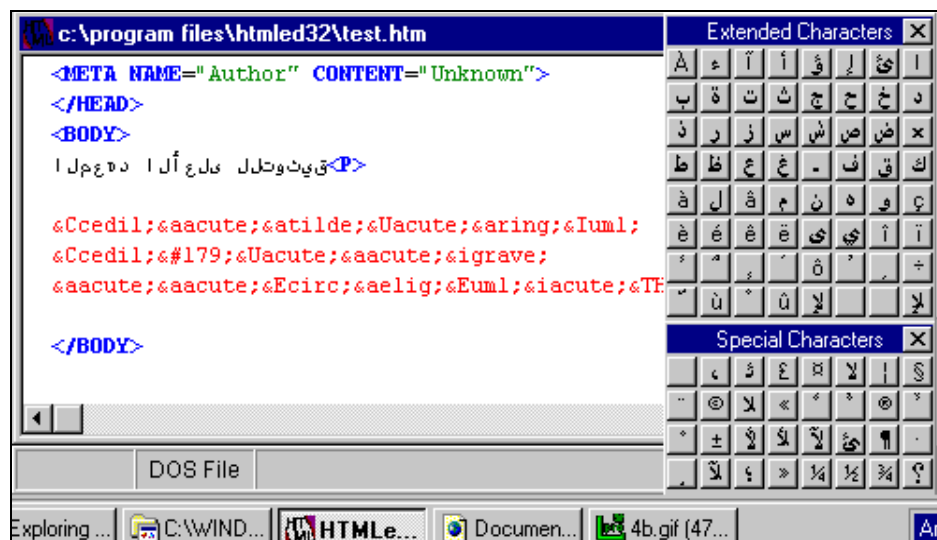


Figure 85 : Caractères arabes en fontes explicites ou par références numériques et alphabétiques

Phase de rendu :

Bien qu'Internet Explorer 3.0 n'ait pas reconnu sous un environnement Windows arabe la forme des caractères arabes dans leurs deux formes de saisie (explicite et implicite), il les affiche par conséquent conformément à leurs équivalents en latin étendu dans la table des codes active.

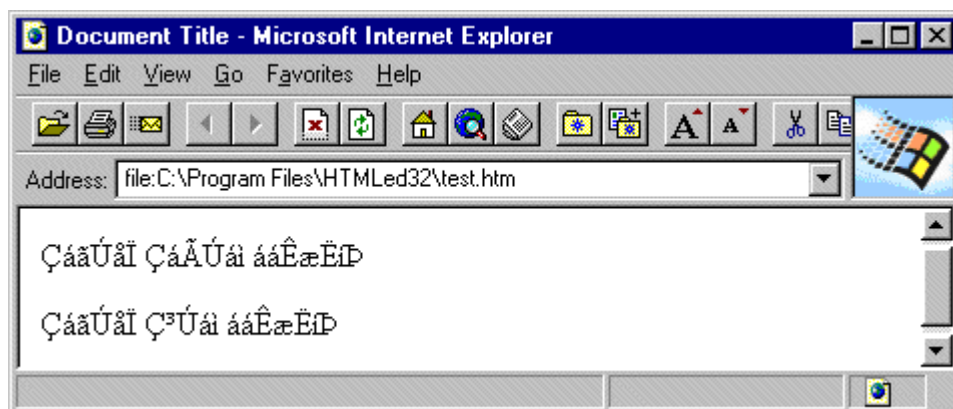


Figure 86 : Caractères arabes non reconnus sous MSIE Explorer 3.0

Test 2 : Caractères arabes dans un éditeur HTML latin et agent utilisateur latin sous Windows arabe AVEC choix explicite par fontes personnalisées

Phase d'édition :

Elle est conforme à la même phase du test précédent.

Phase de paramétrage explicite des fontes à largeur fixe et proportionnelle :

Avant de visualiser le rendu de cette phase d'édition, nous avons procédé au paramétrage de MSIE Explorer 3.0 pour spécifier les fontes personnalisées à largeurs fixe et proportionnelle.

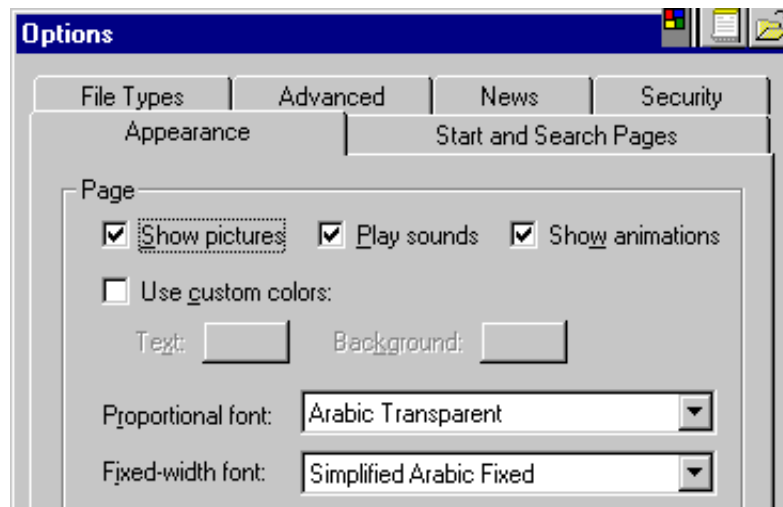


Figure 87 : Paramétrage des fontes arabes à largeur fixe et proportionnelle sous MSIE Explorer 3.0

Phase du rendu paramétré :

Le rendu du texte s'effectuera en tenant compte de ce paramétrage fixé au préalable au niveau de l'agent utilisateur. La restitution s'effectuera ainsi de façon uniforme et adéquate en tenant compte des deux modes de saisie en formats de glyphes réels ou par références alphabétiques et numériques.

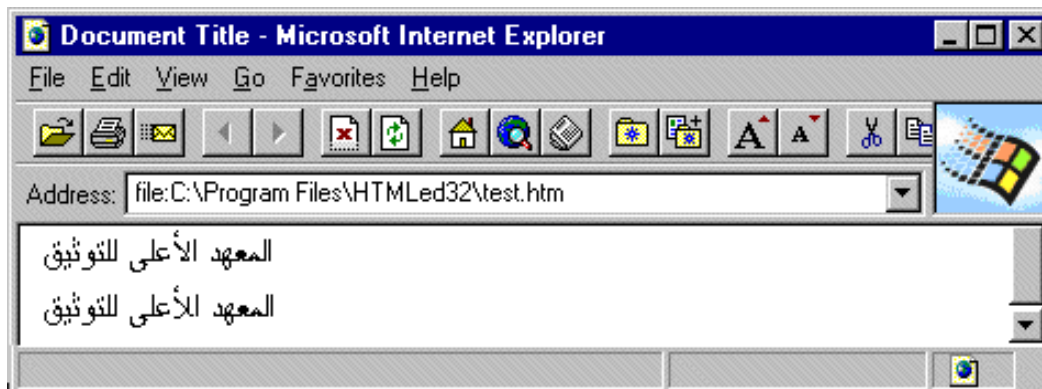


Figure 88 : Reconnaissance uniforme des fontes explicites et implicites

Test 3 : Caractères arabes dans un éditeur HTML latin et agent utilisateur latin sous Windows Arabe AVEC choix explicite de fontes par l'attribut "FACE"

Phase de saisie et d'identification des glyphes arabes réels par l'attribut "FACE" :

Le texte arabe en mode glyphes réels des exemples précédents est identifié par l'attribut "FACE" au sein de l'élément . Au niveau de l'agent utilisateur, les paramétrages des fontes à largeur fixe et proportionnelle est fixé aux valeurs par défaut sur les fontes latines.

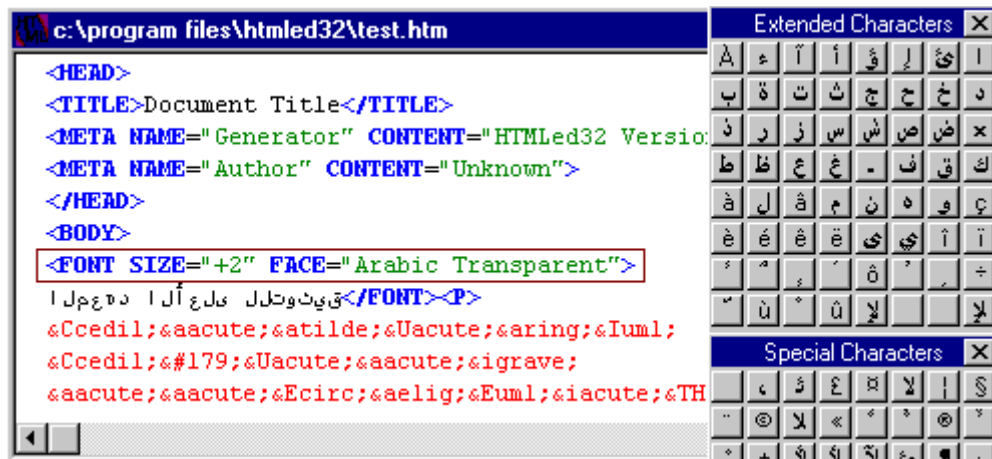


Figure 89 : L'attribut "FACE" ajouté aux glyphes arabes réels par l'élément

Phase de rendu des glyphes arabes identifiés par l'attribut "FACE" :

Le rendu de cette opération visualise les caractères arabes de façon adéquate grâce à l'attribut "FACE" et restitue les références alphanumériques des mêmes caractères en mode Latin-étendu.

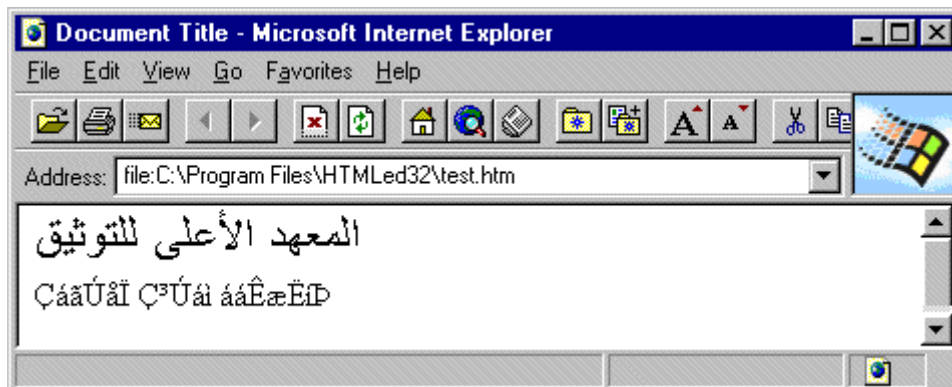


Figure 90 : Effet de l'attribut "FACE" sur les glyphes réels des caractères arabes

Phase de saisie et d'identification des références alphanumériques des caractères arabes par l'attribut "FACE" :

Le texte arabe en mode références alphanumériques de l'exemple précédent est identifié à son tour par l'attribut "FACE" au sein de l'élément . Au niveau de l'agent utilisateur, les paramétrages des fontes à largeurs fixes et proportionnelles restent toujours fixés aux valeurs par défaut sur les fontes latines

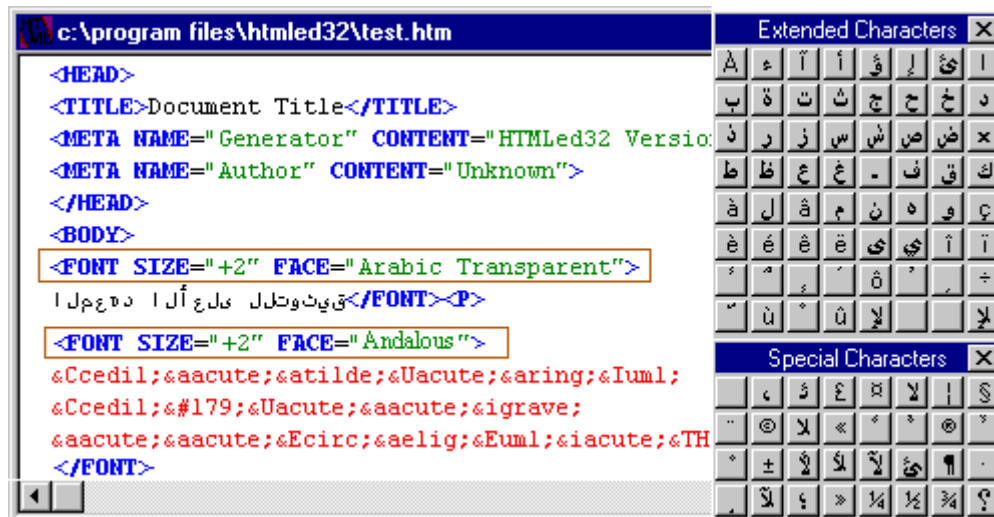


Figure 91 : L'attribut "FACE" ajouté aux glyphes réels et aux références alphanumériques des caractères arabes

Phase de rendu des références alphanumériques des caractères arabes identifiés par l'attribut "FACE" :

Le rendu de cette opération visualise les deux modes de saisie des caractères arabes de façon adéquate grâce à l'attribut "FACE".



Figure 92 : Effet de l'attribut "FACE" sur les glyphes réels et les références alphanumériques des caractères arabes

Test 4 : Caractères arabes dans un éditeur HTML latin et agent utilisateur latin sous Windows Latin AVEC choix explicite par fontes personnalisées

Phase d'édition du texte arabe par références alphanumériques des caractères :

Nous avons repris le même texte des exemples précédents mais dans un contexte purement latin (éditeur, agent utilisateur et plate-forme de travail) afin de vérifier les conditions de faisabilité d'une transparence linguistique arabe universelle et non-propriétaire. L'édition des glyphes réels des caractères arabes étant impossible sur un clavier latin, nous avons procédé à une saisie par références alphanumériques en utilisant la barre d'outils des caractères étendus et spéciaux de HTMLEdit32. Étant donnée que les deux environnements arabe et latin fonctionnent respectivement sous les normes ISO 8859-1 et ISO 8859-6, les deux barres reproduisent respectivement les points de codes au delà des 128 premiers caractères ASCII des deux tables en question.

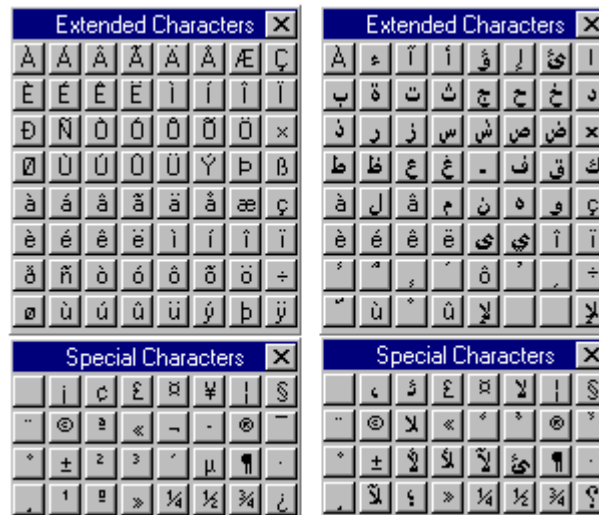


Figure 93 : Les deux tables d'extension de codes selon ISO 8859-1 et ISO 8859-6

La confrontation des deux tables nous a permis de localiser l'emplacement des caractères qui constituent le texte échantillon. Une balise d'alignement droite a été également ajoutée au texte par soucis de conformité avec la directionnalité droite-gauche de l'écriture arabe.

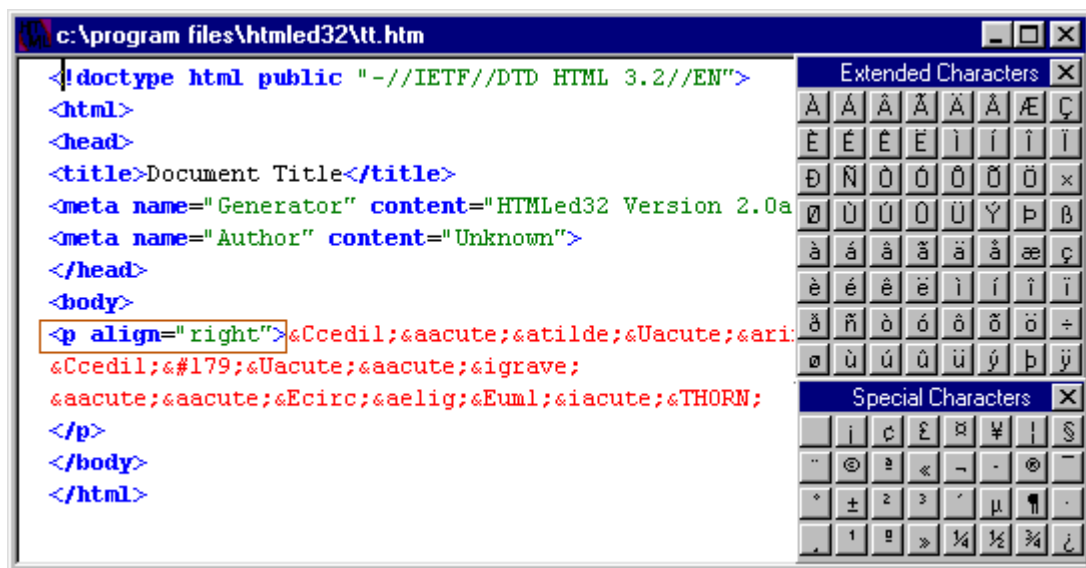


Figure 94 : Édition d'un texte arabe dans un environnement latin intégral par références alphanumériques

Phase de rendu des références alphanumériques des caractères arabes par paramétrage de fontes personnalisées :

Avant la réalisation du rendu suivant, un paramétrage des fontes personnalisées est accompli conformément à la [figure 86](#) précédente. Cette opération est elle-même tributaire d'une opération préalable de déchargement et d'installation d'une fonte arabe appropriée au codage 8 bits conforme à ISO 8859-6. Une fonte arabe Unicode (i.e. Malik Lt BT, Shuweifat) donnerait un résultat différent.

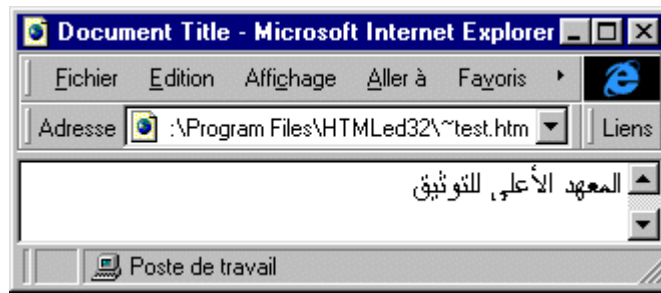


Figure 95 : Rendu des références alphanumériques des caractères arabes justifiés à droite dans un environnement latin intégral par paramétrage de fontes personnalisées

Phase de paramétrage des références alphanumériques des caractères arabes par l'attribut "FACE" :

De la même manière que les tests précédents, la chaîne des caractères arabes saisie par références alphanumériques est paramétrée par l'attribut "FACE" de l'élément pour un rendu arabe dans un environnement latin intégral.

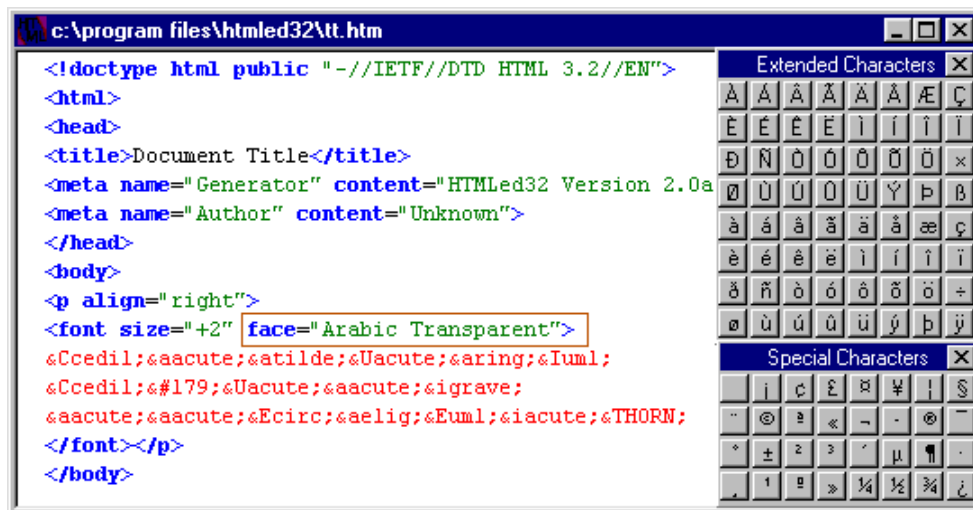


Figure 96 : Paramétrage du rendu d'un texte arabe dans un environnement latin intégral par l'option "FACE"

Phase de rendu des références alphanumériques des caractères arabes sous l'effet de l'attribut "FACE"

Le rendu est conforme aux attentes de l'utilisateur arabophone sur une plate-forme purement latine sans pour autant écarter les difficultés inhérentes à sa mise en œuvre initiale.

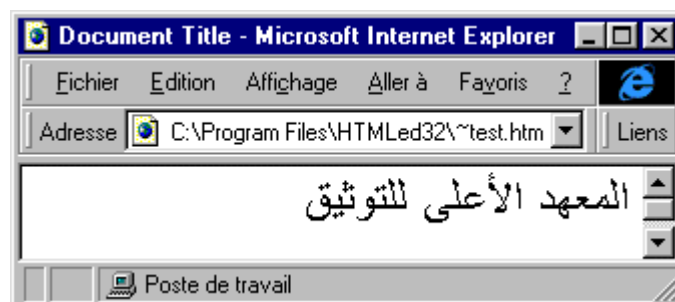


Figure 97 : Rendu arabe dans un environnement latin intégral paramétré par l'option "FACE"

En effet, cette méthode nécessite en particulier le téléchargement d'une fonte appropriée pour qu'elle serve de source d'affichage de la page ou de la séquence codée pour la même fonte. Il est donc important de préciser que cette méthode dispose de deux avantages par rapport aux solutions précédentes :

- ❑ toutes les écritures sont susceptibles d'être intégrées dans le même document à condition de télécharger les fontes qui les concernent. Les caractères latins (par défaut) peuvent ainsi côtoyer les caractères arabes, cyrilliques, hébreux et grecs dans le même document ;
- ❑ l'usage de l'attribut "FACE" conserve en parallèle les valeurs des caractères proportionnels et des caractères à largeurs fixes personnalisés par l'utilisateur.

Certains inconvénients sont pourtant enregistrés à son compte :

- ❑ Le téléchargement des fontes appropriées est réalisé au préalable par l'utilisateur au lieu d'être déclenché automatiquement par l'agent utilisateur au moment de l'accès au document multilingue ;
- ❑ Les écritures à directionnalité Droite-Gauche nécessiteraient un codage supplémentaire pour aligner le texte à droite ;
- ❑ Les utilisateurs des navigateurs autres que Netscape ou Internet Explorer risquent de ne pas interpréter correctement l'option .

g. L'État de l'art

Si toutes ces solutions restent encore à des niveaux très décalés dans la liste des choix des auteurs des ressources multilingues, l'analyse du contexte de l'i18n d'Internet par rapport à la langue arabe fait ressortir les état de faits suivants :

Des solutions modulaires

L'usage de ces caractères arabes dans le contexte Internet fait face à une modularité composée de plates-formes de travail. Les différentes situations dans lesquelles se trouverait un utilisateur potentiel sont les suivantes :

- **Windows arabe / Navigateur Arabe**

Dans ce cas de figure, la problématique est résolue doublement : les deux difficultés de l'arabe en présence du latin, à savoir la directionnalité droite-gauche de l'écriture et la présence des fontes nécessaires pour le rendu du texte arabe sont effectuées par la plate-forme et le navigateur à la fois. La meilleure combinaison actuelle est celle de Windows 95 version arabe et MSI-Explorer 4.01 version arabisée.

- **Windows Arabe / Navigateur Latin**

Ce cas de figure crée un problème d'orientation au niveau de l'interface du navigateur latin qui interprète souvent les appels de directionnalité droite-gauche du système d'exploitation de manière à inverser (ou étirer) les zones des données des documents vers la partie droite de l'écran. Ceci engendre une perturbation dans la disposition des zones de saisie des textes et des boutons cliquables des formulaires. Nous avons observé ce phénomène sur une interface MSIE Explorer 4.0 latine installée sur une plate-forme Windows 95 arabe.

La solution à l'unique problème de la directionnalité que nous avons pu appliquer était de télécharger une version arabisée de MSIE Explorer 4.01

- **Windows Latin / Navigateur arabe**

Sur une plate-forme Windows Latine, la problématique est plus souple. Partant du fait que toutes applications arabisées (y compris les navigateurs Internet), incluent automatiquement une série de fontes arabes, l'affichage sera tributaire du paramétrage adéquat du navigateur pour le choix d'un jeu de caractères codés conforme à la source du document. Il est courant dans ce cas de figure de voir l'agent utilisateur détecter automatiquement l'encodage utilisé par la ressource incluse dans le paramètre "CHARSET" de l'en-tête du document. Comme le démontre la figure suivante, un changement de codes peut toujours influencer le rendu du texte y compris dans la même langue.



Figure 98 : Navigateur multilingue sur plate-forme latine

- **Windows Latin / Navigateur Latin**

C'est la configuration la plus problématique car elle nécessite un forçage double de la reconnaissance des glyphes arabes et de la bidirectionnalité Gauche-Droite et Droite-Gauche de l'écriture bilingue. C'est également la configuration la plus urgente à développer pour contourner les deux contraintes suivantes :

- ❑ La limite de l'étendue géographique de la diffusion et de l'accès aux ressources multilingues arabes-latines ;
- ❑ L'obligation de faire usage des solutions propriétaires du domaine commercial pour accéder et diffuser une information scientifique et technique multilingue.

À priori, et dans l'attente du développement des plates-formes et des applications entièrement Unicode, des solutions d'adaptation assez complexes sont possibles pour forcer, par un paramétrage manuel, l'affichage des caractères arabes dans une forme et une directionnalité adéquates au contexte linguistique arabe. Ceci nécessiterait le passage par une série de procédures à accomplir par un enchaînement aléatoire :

La première initiative consiste à télécharger et installer des fontes arabes adéquates sur le système latin. Bien que le domaine des fontes fasse l'objet d'une activité commerciale abondante, certaines d'entre elles sont fournies à titre gratuit.

La deuxième étape consiste à utiliser les fontes arabes comme paramètre de définition des fontes personnalisées à largeurs fixes et variables (fig.80, fig.81, fig.87). Il est normalement plus approprié que l'agent utilisateur le fasse automatiquement pour l'utilisateur selon les principes de la négociation linguistique que nous avons analysée auparavant ; mais étant donnée que la plateforme de la configuration que nous analysons est par défaut latine, l'utilisateur est censé accomplir ce paramétrage de façon manuelle (fig.46, et fig.47).

Des solutions intégrées

Face à la précarité de ces solutions intermédiaires, des produits intégrés d'Internet sur la base d'une reconnaissance du caractère et de la langue arabe sont apparues sur le marché pour proposer des solutions intégrales de saisie, de traitement, de diffusion et d'accès aux ressources multilingues. Nous en retenons quelques-uns dont un descriptif sommaire figure dans la première partie de ce travail (§ I-6.2).

Le tableau suivant décrit sommairement la portée de quelques-uns des navigateurs Internet les plus connus dans le support de la langue arabe. La plus importante observation à formuler concernant ce tableau est l'importance seconde que toutes ces solutions ont donnée au critère de la directionnalité Droite-Gauche et à l'analyse contextuelle des caractères arabes.

Comparaison	Tango	Arabic Internet Explorer	Netscape	Sindbad Plug-in	Internet with an Accent
Concepteur	Alis Technologies Inc.	Microsoft Co.	Netscape Communications Co.	Sakhr	AccentSoft
Version du navigateur	3.0	3.0	3.0	Plug-in (Non pas le navigateur)	1.0
Plates-formes	Toutes versions Windows	Arabic Windows seulement	Arabic Windows seulement	Arabic Windows seulement	Toutes versions Windows
Interface Arabe	Oui	Non	Non	Oui	Oui
Changement automatique de l'interface	Oui	Non	Non	Non	Oui
Justification à droite de l'Interface arabe	Oui	Non appliqué	Non appliqué	Oui	Non
Justification Droite	Oui	Oui	Non appliqué	Non	Non
Mode assistance arabe en ligne	Oui	Non	Non	Non	Non
Préférences linguistiques	90	Non	Oui	Non	1
Boutons linguistiques	Oui	Non	Non	Non	Non
Claviers sur écran	Oui	Non	Non	Non	Non
Support de la directionnalité Droite-Gauche	Oui	Non	Non	Partial	Non
Support du codage HTML pour la langue arabe	Oui	Oui	Non	Non	Oui
Support des pages de codes arabes	6	1	1	13	3
Messagerie arabe	Oui	Non	Non	Oui (Arabic Windows seulement)	Non

Tableau 30 : Degrés de traitement de la langue arabe par les navigateurs Internet Multilingues

Source : Traduit et adapté de chez Alis technologies Inc. <http://www.alis.com>. Dernière modification : 5 janvier 1998

Dans sa version 4.01 avec support arabe, Microsoft Internet Explorer fait un gain de surface dans le contexte de l'i18n dans le monde arabe. La solution reste cependant encore limitative tant la transparence linguistique n'est pas encore devenue à la portée de toutes les plates-formes sans distinctions géographiques ni linguistiques.

4.2.2.4 CONCLUSION

En définitive, devant la multitude de solutions mises en œuvre pour traiter le phénomène de l'i18n de façon totale ou partielle, il est communément recommandé que les agents utilisateurs doivent observer l'ordre décroissant des priorités suivantes chaque fois qu'ils sont appelés à déterminer l'encodage des caractères d'un document.

1. Une action implicite de l'utilisateur pour résoudre un désordre quelconque dans le rendu d'un document quelconque (i.e. choix manuel d'un jeu de caractères codés) ;
2. Le paramètre HTTP « Charset » dans le champ « Content-Type » ;
3. La déclaration META à travers la valeur de « http-equiv » définie pour désigner le paramètre « Charset » dans le champ « Content-Type » ;
4. Le paramètre « Charset » attribué aux éléments A & LINK ;
5. Les heuristiques de l'agent utilisateur et les paramètres personnels de l'utilisateur lui-même. Certains agents utilisateurs, en l'absence d'une indication d'un jeu de caractère pour le document en cours d'utilisation, appliquent leurs propres jeux des caractères codés par défaut (i.e. ISO 8859-1).

4.2.3 Les apports HTML au profit du multilinguisme d'Internet

Les standards de l'i18n de HTML sont aujourd'hui en pleine effervescence bien que leur application souffre encore des insuffisances de conformité et d'unanimité d'usage. Comme le signale G.Nicol, « *WWW browsers available today are uniformly incomplete in their support of input and display methods for dates, time, weight, and other such things, in a manner that seems natural to the end user. Even localised products fail here. The primary reason is that there is no strictly defined way of representing such information in the WWW. Rather, it is left to the authors' discretion* »^[148].

Dans sa version actuelle, HTML 4.0 fournit une base assez importante d'éléments associés à l'internationalisation du Web parmi lesquels certains spécifient la langue du document intégral ou des parties qui le constituent, les listes des items qu'il comprend et les en-têtes qu'il contient. Une fois le navigateur déchiffre la langue du document, il déploie implicitement un mécanisme de reconnaissance automatique des formes appropriées à certains caractères de ponctuation, d'union et de ligature. HTML 4.0 gère également la bidirectionnalité des textes multilingues arabe/latins par un balisage particulier.

[148] Gavin T. NICOL. The Multilingual World Wide Web. <http://www.sil.org/sgml/nicol-multwww.html>

4.2.3.1 Un brin d'histoire

Depuis sa création par Tim Berners-Lee dans les laboratoires du CERN^[149], et son expansion grâce au navigateur Mosaic de la NCSA^[150], HTML a connu une amélioration rapide pour devenir la plate-forme par défaut de toutes les applications et les protocoles Internet^[151]. Son évolution, comme tout produit ou service de ce genre, est définie par les RFCs qui consolident progressivement sa portabilité multiplates-formes et l'unanimité de son usage universel. À partir de la version HTML 2.0 (RFC 1866), l'IETF (Internet Engineering Task Force) prend la charge du développement de la norme en question sur le principe de la concertation internationale afin d'asseoir une uniformité d'usage des spécificités de codage HTML. La version 3.0 (1995) était une version améliorée de HTML 2.0 bien qu'il n'ait pas atteint un grand taux d'unanimité d'usage. Les efforts du groupe de travail^[152] du Consortium WWW (W3C)^[153] pour un usage uniforme d'HTML dans toutes leurs activités ont donné lieu à la version HTML 3.2 en janvier 1997. L'objectif de ces mises à jours était essentiellement d'étendre les spécificités du codage HTML pour garantir l'interopérabilité des divers navigateurs sur diverses plates-formes et éviter ainsi que le Web ne se décompose en sous systèmes de formats propriétaires, indépendants et incompatibles. Les dernières évolutions de HTML étaient ainsi guidées par une vision d'ouverture et d'interopérabilité totale des systèmes. La version HTML 4.0 constitue actuellement l'initiative première vers le nouveau paradigme de l'i18n et de la l10n d'Internet. Elle enregistre ainsi une démarcation importante par rapport à ses prédécesseurs sur l'aspect multilingue des ressources et services Internet. En fait, HTML 4.0 étend la portée du codage HTML vers d'autres spécificités techniques importantes : usage des feuilles de style ([CSS](#)), amélioration des techniques des cadres (frames), des formulaires (forms) et des objets incorporés (HREF), amélioration de l'accessibilité aux handicapés et ... amélioration du support multilingue et du traitement bidirectionnel des textes.

Nous mettrons l'accent sur ces deux derniers critères, en tant que valeurs ajoutées apportées à la dite version par la contribution d'experts en i18n^[154], pour voir en profondeur les apports réels de HTML 4.0 dans le traitement des documents multilingues et évaluer ces apports en fonction des particularités du contexte arabe/latin que nous avons soulevées jusque là.

Avant d'aborder l'aspect multilingue de HTML, il nous semble important de soulever le point fondamental concernant les origines de HTML et ses sources d'inspiration dans son prédécesseur SGML. L'influence de ce dernier est plus qu'évidente dans la portée multilingue du premier.

4.2.3.2 HTML 4.0 / SGML, entre la structure et le style des documents électroniques

HTML puise ses origines dans le langage SGML, considéré comme un langage de balisage pour l'organisation des champs sémantiques dans les documents structurés. Il constitue l'une des formes d'application de la norme ISO 8879 datée de 1986^[155]. Cathérine LELOU précise que « *HTML est une déclinaison de la norme SGML, avec une DTD simplissime* »^[156].

[149] En mars 1989, Tim BERNERS-LEE du CERN publia l'article « Hypertexte et le CERN ».

[150] En février 1993, Marc ANDREESEN édite la première version du browser Mosaic sous Xwindows par la NCSA.

[151] Le principe de HTML est repris de la technique Hypertexte inventée à l'origine par Ted Nelson en 1965.

[152] Le *W3C HTML Working Group* est composé des sociétés Adobe, Hewlett Packard, IBM, Microsoft, Netscape Communications Corporation, Novell, SoftQuad, and Sun Microsystems.

[153] Créé en juillet 1993 par le Cern et le MIT puis l'INRIA, le WWW Consortium œuvre pour la normalisation du Web.

[154] About the HTML 4.0 specification. W3C Working Draft 17-Sep-1997. <http://www.w3.org/TR/WD-html40-970917/>

[155] ISO 8879. Information Processing Text and Office Systems — Standard Generalized Markup Language (SGML). 1986.

[156] Cathérine LELOU. Moteurs d'indexation et de recherche. Op. Cit. p.43.

Avec le développement de HTML, ses éléments et attributs de présentation sont périodiquement remplacés par des mécanismes novateurs qui lui permettent de gagner en qualité et en maturité. L'évolution des techniques de codage et de balisage vers HTML 4.0 a engendré le développement de plusieurs autres méthodes d'étiquetage et de négociation linguistique des documents HTML, empruntées entre autres à SGML et à MIME. L'imbrication des Documents des Types de Données (Data Type Documents) généralement connues sous l'acronyme DTD et des feuilles de styles (Cascading Style Sheets) mieux connues sous l'acronyme de [CSS](#), constituent des exemples concrets de l'enrichissement constant de HTML. Une DTD sous HTML est à titre d'exemple une forme officielle de définition de la syntaxe HTML dans des termes propres à SGML.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
"http://www.w3.org/TR/REC-html40/loose.dtd">
<HTML>
<HEAD>
  <META http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
  <TITLE>W3C Internationalization and Localization</TITLE>
  <LINK rel=stylesheet href="i18n.css">
</HEAD>
<BODY BGCOLOR="#ffffff" TEXT="#000000">
...
</BODY>
```

Figure 99 : En-tête HTTP enrichie par les éléments DTD et CSS

Pour rappeler la structure HTML à travers cet exemple, un document HTML 4.0 est généralement composé de trois parties :

1. Une partie qui indique la version HTML utilisée ;

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
"http://www.w3.org/TR/REC-html40/loose.dtd">
```

Un document HTML valide déclare la version HTML utilisée dans le document courant. Cette partie inclut également la nature de la DTD (*Document Type Declaration*) utilisée dans le document courant.

HTML 4.0 spécifie trois types de DTD :

1 - Une DTD Stricte : inclut tous les éléments et les attributs qui n'ont pas été déclarés obsolètes ou qui n'apparaissent pas dans des documents à bases de cadres (frames). HTML 4.0 recommande d'utiliser ce type de DTD dans la forme de déclaration suivante :

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
"http://www.w3.org/TR/REC-html40/strict.dtd">
```

2 - Une DTD transitionnelle : inclut les paramètres de la DTD Stricte auxquels elle ajoute les éléments et les attributs de HTML déclarés obsolètes. HTML 4.0 recommande d'utiliser ce type de DTD dans la forme de déclaration suivante :

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
"http://www.w3.org/TR/REC-html40/loose.dtd">
```

3 - Une DTD pour cadres (frameset) : inclut tous les paramètres de la DTD transitionnelle auxquels elle rajoute les éléments et les attributs des cadres. HTML 4.0 recommande d'utiliser ce type de DTD dans la forme de déclaration suivante :

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
"http://www.w3.org/TR/REC-html40/frameset.dtd">
```

Le consortium W3C recommande l'usage des DTD strictes qui excluent les éléments et les attributs de présentation désormais confiés aux techniques des feuilles de style.

2. Une section d'en-tête déclarative ;

```
<HEAD>
  <META http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
  <TITLE>W3C Internationalization and Localization</TITLE>
  <LINK rel=stylesheet href="i18n.css">
</HEAD>
```

3. Un corps de document qui structure le contenu intellectuel du document ;

```
<BODY BGCOLOR="#ffffff" TEXT="#000000">
...
</BODY>
```

Bien qu'ils constituent des additifs au delà des exigences linguistiques pures, les paramètres DTD de SGML et les feuilles de style traitent entre autres des solutions plus sophistiquées de négociation et d'étiquetage linguistique.

4.2.3.2.1 SGML : une flexibilité de balisage

SGML est un métalangage pour la définition des formes de balisage des textes que les auteurs utilisent pour représenter l'information structurelle, sémantique et de formatage de leurs documents. HTML n'en est qu'un exemple d'application.

Étant donné qu'il y a un grand nombre de techniques d'encodage de texte, SGML ne s'adresse pas directement à la manière comment une chaîne de caractères dans un document SGML est représentée par une série d'octets. Cet encodage est considéré dans la dimension du document SGML comme un encodage externe des caractères qui diffère substantiellement du jeu de caractères codés dans les documents HTML. En effet, SGML traite les jeux des caractères comme un répertoire de caractères représentés par un jeu de codes qui attribuent à chaque caractère un identifiant unique.

Gavin T. Nicole définit SGML dans les termes suivants :

"SGML does not define the representation of characters. ISO 8879:1986 defines a code set as a set of bit combinations of equal size, a character as an atom of information, a character repertoire as a group of characters, and a character set as a mapping of a character repertoire onto a code set such that each character is represented by a unique bit combination within the code set. An SGML parser deals with characters, as defined above, and hence, is logically independent of the physical representation of the data. There

is often an internal representation of characters that could be quite different to that used in data storage^[157].

Par contre, HTML en tant qu'une forme d'application de SGML, ne s'adresse pas directement à la question de l'encodage des caractères externes. Ceci est laissé aux soins des mécanismes externes comme MIME.

Parmi les autres caractéristiques de SGML qui pourraient nous intéresser à ce niveau d'analyse, la notion de balisage libre est un point important dans la structuration-auteur des documents et de leurs méthodes de rendus paramétrables par l'utilisateur. Le simple recours à SGML en tant que langage généralisé de balisage des documents structurés est à notre avis un gain de qualité pour le réseau Internet. Notre conviction de cette proposition émane des faits essentiels suivants :

- ❑ SGML est un langage suffisamment riche et flexible pour fournir le maximum de solutions pour toutes les situations et les besoins de balisage des documents. Il ouvre ainsi l'univers Web à des ressources qui dépassent parfois les limites du langage HTML ;
- ❑ SGML permet de publier toute information conforme aux attentes des lecteurs et non pas uniquement ce que le langage de balisage permet de faire ;
- ❑ SGML permet une méthode de recherche plus performante grâce à sa méthode de structuration sémantique ;
- ❑ SGML fournit une solution non propriétaire du domaine public pour l'échange de toutes sortes de données.

La jonction entre les documents SGML et les navigateurs HTML s'effectuait par conversion des documents d'un format à un autre. Les navigateurs WWW s'accommodent davantage à cette forme de documents SGML par la méthode des Plug-In sous forme d'un navigateur (parser) SGML qui se déclenche automatiquement, à l'instar d'une visionneuse (viewer) graphique, chaque fois qu'il rencontre un fichier avec une extension « sgm » ou « sgml ». Dans cette catégorie des navigateurs SGML indépendants, la société [Softquad](#)^[158], a mis en place un navigateur du domaine public pour l'affichage des documents SGML, en l'occurrence le navigateur Panorama^[159]. Cette version nécessite cependant le téléchargement de la fonte qui sera utilisée avec l'utilitaire en question.

Une meilleure solution serait toutefois de créer l'extension interne des agents utilisateurs afin de visualiser les documents SGML en mode standard. Le développement de cette approche serait rendu possible avec l'enrichissement des potentialités actuelles et futures des définitions des types de données (DTD) et des feuilles de style (CSS) qui disposent déjà de certaines primitives de rendu de chaque type d'élément.

Les DTD et les CSS peuvent, en effet, être conçues de façon à donner à l'auteur des ressources Internet une marge de flexibilité dans la structuration de ses documents sans besoin de paramétrage particulier de l'agent utilisateur et sans porter préjudice à la performance de rendu du document en question. Pour certains types de DTD (surtout celles de HTML), les feuilles de style sont distribuées avec le navigateur. Dans d'autres cas, le navigateur ira les chercher dans le serveur ([fig.98](#)).

[157] Gavin T. NICOL. The Multilingual World Wide Web. Op.Cit.

[158] SofQuad Inc. <http://www.sq.com/>

[159] Panorama est conçu en plusieurs versions : SoftQuad Panorama, SoftQuad Panorama PRO et SoftQuad HoTMetal PRO; tous sont des marques déposées de SoftQuad Inc. URL : <http://www.sq.com/>

Dans la génération des agents utilisateurs actuels, la façon de traiter les balises de formatage est strictement définie par les performances de l'agent utilisateur : au moment de la rédaction des codes de formatage ou de leur restitution, l'auteur et l'utilisateur restent généralement tributaires de la nature des attributs utilisés conformément à la version HTML en cours. Bien que ceci n'est pas toujours vrai (i.e. les cas des attributs des fontes modifiables à la guise de l'utilisateur), la marge des balises et de leurs attributs restent fixée par les spécificités de la version HTML utilisée.

Notre proposition à l'égard de cet état de fait consiste à créer davantage de flexibilité des documents HTML et SGML par le développement d'une dimension plus dynamique dans les phases de la conception et de la restitution des ressources HTML. Des règles de traitement des balises pourraient être automatiquement chargées avec le document restitué afin de permettre son traitement personnalisé. Une table contenant toutes les balises dans le document accompagnerait dans ce cas concret le document restitué pour renseigner sur les options choisies par l'auteur de la ressource.

Deux pré-requis essentiels sont indispensables pour atteindre cet objectif :

1. Chaque document HTML devrait avoir, depuis sa création, une liste exhaustive des balises utilisées et une table de description des méthodes de leur traitement (attributs). Ceci pourrait prendre forme d'une déclaration DTD dans l'en-tête du document.
2. Des codes d'extension pour le traitement des balises aléatoires non déclarées dans la liste associée.

Parmi les retombées escomptées à cette approche, l'indépendance de l'auteur des ressources vis à vis des contraintes des agents utilisateurs (i.e. les cas des contraintes des frames avec les agents utilisateurs anciens) constitue un gain de qualité et de transparence remarquable. L'objectif définitif d'une accentuation des performances des outils et des ressources Internet devrait à notre sens aboutir à une "indépendantisation" de l'utilisateur et du concepteur des langages de balisage et de leurs extensions aléatoires et limitatives et laisser la force de la sémantique au bout des doigts des auteurs et non des programmeurs. Les feuilles de style dans leur version 2 jouerait certainement un grand rôle dans l'aboutissement à cette flexibilité sémantique et linguistique des ressources Internet.

4.2.3.2 Les feuilles de Style : un apport à l'i18n

Le domaine des feuilles de style est l'un des axes sur lesquels le W3C œuvre de plein pied à travers son *Internationalization Working Group* pour renforcer les initiatives de l'i18n du Web^[160].

Le paramétrage des propriétés des fontes est l'une des grandes priorités dans la conception des feuilles de style. Étant donnée qu'il y a un problème évident au niveau de la taxonomie de la classification des fontes (« Italique » est couramment utilisé pour identifier le texte incliné qui peut l'être aussi par l'identifiant « Oblique » ou « Cursive » ou « Incliné »), il n'est généralement pas si évident de faire la correspondance entre les propriétés d'une fonte abstraite et sa valeur.

Les documents multilingues nécessitent un support de fontes pour afficher des caractères non disponibles au niveau des clients. Un caractère arabe ne serait pas implicitement affiché sous un navigateur latin, et un caractère latin ne le serait pas non plus sous un navigateur cyrillique sans

[160] W3C activity: Internationalization. <http://www.w3.org/International/Activity.html>

la présence d'une fonte appropriée. Le problème à ce niveau est la grande diversité des caractères de toutes les langues du monde. Chaque caractère nécessiterait alors au moins un glyphe pour être affiché convenablement.

L'intérêt porté aux propriétés de style pour résoudre cet aspect de traitement des documents multilingues, émane de l'intérêt croissant que revêt cette nouvelle méthode dans la désignation des valeurs linguistiques d'un document électronique par le traitement implicite des caractères qui le constituent. C'est l'un des moyens actuellement en vogue pour contourner la rigidité des navigateurs Internet classiques dans le traitement des documents multilingues. Les propriétés de style couvrent une grande panoplie de règles et de mécanismes de traitement implicite des documents HTML. Elles peuvent être imbriquées dans des éléments particuliers du balisage HTML, regroupées au niveau de la partie en-tête du document ou placées séparément dans une feuille de style associée. Nous nous concentrerons dans notre étude sur l'aspect linguistique et plus particulièrement le traitement des glyphes, des fontes et des caractères.

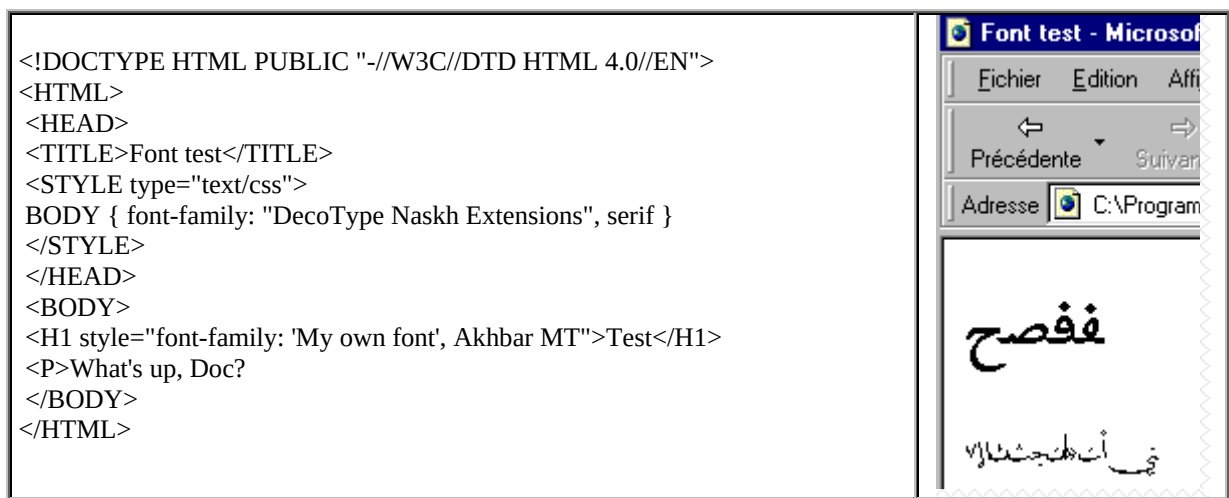


Figure 100: Traitement linguistique par règles de style au sein de l'en-tête et du corps d'un document HTML

Pour contourner le problème de la seule fonte qui ne pourrait pas représenter tous les caractères dans un document, les règles de style permettent d'utiliser un jeu de fontes d'un même style et d'une même taille dans lequel le système cherchera l'existence d'un glyphe correspondant à un caractère particulier.

La ligne **"BODY { font-family: "DecoType Naskh Extensions", serif }**" dans la figure précédente, indique qu'il est prévu que le texte dans le corps du document contienne du texte en caractères arabes pris dans la fonte "Deco Type Naskh extension" et que tout autre caractère est censé être choisi dans la famille de fonte générique "Serif". La famille "Serif" serait également utilisée si la fonte précédente n'est pas disponible.

Quand un document textuel est visualisé par un agent utilisateur, les caractères qui le constituent sont normalement traduits en glyphes. Un ou plusieurs caractères peuvent toutefois correspondre à un ou plusieurs glyphes à la fois (§ IV-1). Pour avancer des définitions succinctes des concepts des glyphes, des fontes et des caractères (ces définitions seront reprises en détail dans la partie suivante de ce travail), nous nous contenterons de dire que le glyphe est la forme de représentation d'un glyphe abstrait qui prend alors forme artistique dans un style typographique visuel sur écran ou sur un support imprimé. Une fonte est un ensemble de glyphes unis par une

forme identique de motif, de taille et d'apparence associée à l'ensemble entier de glyphes. Elle représente l'étape intermédiaire dans le chaînage caractère/glyphe.

Dans toute opération de rendu, un agent utilisateur doit vérifier les faits suivants :

- ❑ S'il existe de façon directe ou indirecte une fonte particulière à un caractère particulier;
- ❑ Si l'agent utilisateur dispose de cette fonte ;
- ❑ Dans l'affirmative à quel(s) glyphe(s) ce(s) caractère(s) correspond(ent) ils ;
- ❑ Dans la négative, comment procéder ? Faudrait-il opter pour une autre fonte ? La fonte pourrait-elle être générée ou pourrait-elle être déchargée d'un serveur Web ?

Toutes ces opérations de contrôle et de résolution des alternatives sont désormais confiées aux feuilles de style dans leurs deux versions CSS1 et CSS2 à travers lesquelles les auteurs spécifient les caractéristiques des fontes par une série de propriétés. L'amélioration de CSS1 vers CSS2 réside essentiellement dans les deux points suivants :

a. Attributs des fontes

Pour ne plus se limiter aux noms des fontes assumées présentes sur le système client, CSS2 spécifie les fontes existantes ainsi que leurs alternatives par leurs attributs ou caractéristiques suivantes :

Font family

C'est le nom du groupe de fontes conçues pour être utilisées ensemble sur le principe de la similitude artistique. Les familles de fontes ne sont pas réservées aux caractères latins. Elles peuvent être constituées sur la base de plusieurs critères comme le fait d'être un caractère avec ou sans sérif ; avoir un espacement proportionnel, être des caractères fantaisistes etc.

Font style

Cette caractéristique spécifie si le caractère doit apparaître dans un style italique ou normal.

Font variant

Cette propriété indique la fonte appropriée d'un caractère et indique le choix entre deux formes de glyphes pour une même fonte.

Font weight

Le poids de la fonte renvoie à la grosseur ou minceur du glyphe utilisé pour le rendu du texte par comparaison à d'autres fontes dans la même famille.

Font stretch

L'espacement des fontes indique les taux de condensation ou d'aération entre les glyphes utilisés pour le rendu du texte par comparaison à d'autres fontes dans la même famille.

Font size

Ce critère indique la taille de la fonte utilisée.

b. Font matching

Avec CSS2, les auteurs de feuilles de styles sont désormais libres de choisir les fontes de leur choix. L'agent utilisateur saura les substituer par des alternatives qui s'y rapprochent sur le plan stylistique le cas où elles ne figurent pas sur le système client. Par l'intermédiaire d'un "algorithme de correspondance de fontes" (*font matching algorithm*), les fontes sont sélectionnées par un rapprochement stylistique à travers leurs attributs et sélectionnées par les quatre méthodes suivantes :

Concordance du nom de fonte (*font name matching*)

L'agent utilisateur utilise dans ce cas une fonte existante qui porte le même nom que celle réclamée par l'auteur du document.

Concordance intelligente des fontes (*intelligent font matching*)

Dans l'absence d'une fonte identique aux spécificités réclamés par l'auteur, l'agent utilisateur invoque une fonte existante et accessible sur le principe du rapprochement stylistique.

Synthèse de fontes (*font synthesis*)

Dans ce cas de figure, l'agent utilisateur crée une nouvelle fonte qui se rapproche le plus des spécificités de la fonte requise en utilisant ses mêmes valeurs métriques (texte ou symbole, nature des serifs, poids, tailles, etc.).

Déchargement de fontes (*font download*)

L'agent utilisateur peut toutefois télécharger une fonte à partir d'un serveur sur le Web à la manière du téléchargement d'une image, d'un son ou d'une applettes.

Cette nouvelle performance est connue sous le nom de "WebFonts". Les "WebFonts" peuvent être utilisés dans les feuilles de style CSS2 ou dans d'autres types de media comme les applettes Java, les scènes VRML ou CGM en ligne.

4.2.3.3 Les perspectives de l'i18n de HTML

L'évolution de HTML depuis sa version initiale n'a cessé d'apporter des modifications au profit de l'i18n du traitement des données. Certaines de ces modifications gardent encore leurs valeurs initiales même avec les versions les plus récentes du standard en question.

Parmi ces modifications enregistrées depuis 1996 sur les tables de HTML, les deux attributs "DIR" et "LANG" ont été ajoutés pour enrichir la portée multilingue de tous les éléments HTML ; les attributs "ID" et "CLAS" renforcent les feuilles de style pour des besoins de structuration.

```
<!ENTITY % attrs
"id ID #IMPLIED -- element identifier --
class NAMES #IMPLIED -- for subclassing elements --
lang NAME #IMPLIED -- as per RFC 1766 --
dir (ltr|rtl) #IMPLIED -- I18N text direction --">
```

4.2.3.3.1 L'attribut « LANG » pour le traitement du langage naturel

La valeur de l'attribut « LANG » est un code de langue spécifié par le RFC 1766 qui identifie une langue naturelle parlée, ou écrite utilisée à des fins de communication humaine de l'information ; raison pour laquelle les langages informatiques sont exclus de façon explicite des codes de langues dans l'attribut « LANG ».

Les codes de langue se présentent sous forme d'un code de base et quand nécessaires d'une série de sous codes disposés selon la forme suivante :

langue primaire-sous code de langue secondaire

Les deux caractères des codes primaires de langues sont définis par la norme ISO 639 (i.e. fr (français), de (allemand), ar (arabe), he (hébreu), ru (russe), zh (Chinois), ja (Japonais), hi (hindi), ur (urdu), sa (sanskrit)... Toute autre paire de lettres pour les sous codes sont définies par la norme ISO 3166. Parmi les codes des langues secondaires on fait souvent référence aux variantes de la langue anglaise spécifiées comme suit :

- ❑ "en" : Anglais
- ❑ "en-US" : L'anglais américain.
- ❑ "en-cockney" : l'anglais Cockney.
- ❑ "i-choke" : la langue Cherokee parlée par certains amérindiens.

Dans l'exemple suivant, la langue primaire d'un document est le français « fr » puisque déclarée à l'en-tête du document comme attribut de l'élément « HTML ». Dans la partie du corps du document, un premier paragraphe est déclaré être en espagnol « es », suivi d'un retour à la langue de base « fr ». Un paragraphe suivant en français comprend une incise arabe déclarée par la paire , "Embedding" à la suite de laquelle, la langue de base est réactivée.

```
<HTML lang="fr">
<BODY>
...Interprété en français...
<P lang="es">...Interprété en espagnol...
<P>...Interprété de nouveau en français...
<P>...Paragraphe français interrompu par <EM lang="ar">du
texte arabe</EM>le français est repris à c e niveau...
</BODY>
</HTML>
```

L'utilité de l'attribut « LANG » se fait observer sur plusieurs aspects du rendu linguistique des documents structurés. Une première utilité concerne la « désambiguïsation » des caractères dans les cas où le simple décodage des caractères serait insuffisant pour identifier le glyphe du caractère correspondant. Nous avons déjà évoqué dans un cas précédent que la valeur hexadécimale %E9 (en format %HH) peut à la fois correspondre au caractère latin « é » ou au caractère cyrillique « IOU » ou à beaucoup d'autres caractères dans d'autres écritures. L'attribut "LANG" permettrait au système décodeur de choisir le glyphe correspondant à la langue désignée. Le même cas se présente dans le traitement des langues arabe et persane qui, si elles se ressemblent sur le plan des glyphes des caractères, diffèrent sur certains points de la ligature, des diacritiques et des caractères composés. L'utilité de l'attribut « LANG » est également vérifiée pour résoudre des problématiques de rendu des marques de ponctuation (i.e. les guillemets

français et anglais), les traits d'union, la ligature, la gestion des espaces et la synthèse de la parole.

L'importance de l'attribut « LANG » en tant que marqueur de contenu linguistique atteint son paroxysme avec les formulaires de requêtes et les moteurs de recherche multilingues.

Quand plusieurs éléments META (i.e. « Keywords ») décrivent le même documents sur la base d'une identification linguistique diversifiée (i.e. mots-clés en plusieurs langues), les moteurs de recherche peuvent filtrer la version correspondante au choix linguistique de l'utilisateur sur la base de l'attribut « LANG ». L'exemple suivant démontre comment indexer par valeurs d'éléments META un seul document par plusieurs versions linguistiques des mêmes concepts. Une fois récupérés par les robots indexeurs dans les index et les répertoires des moteurs de recherche, les documents sont restitués sans considération de la version linguistique originale du documents.

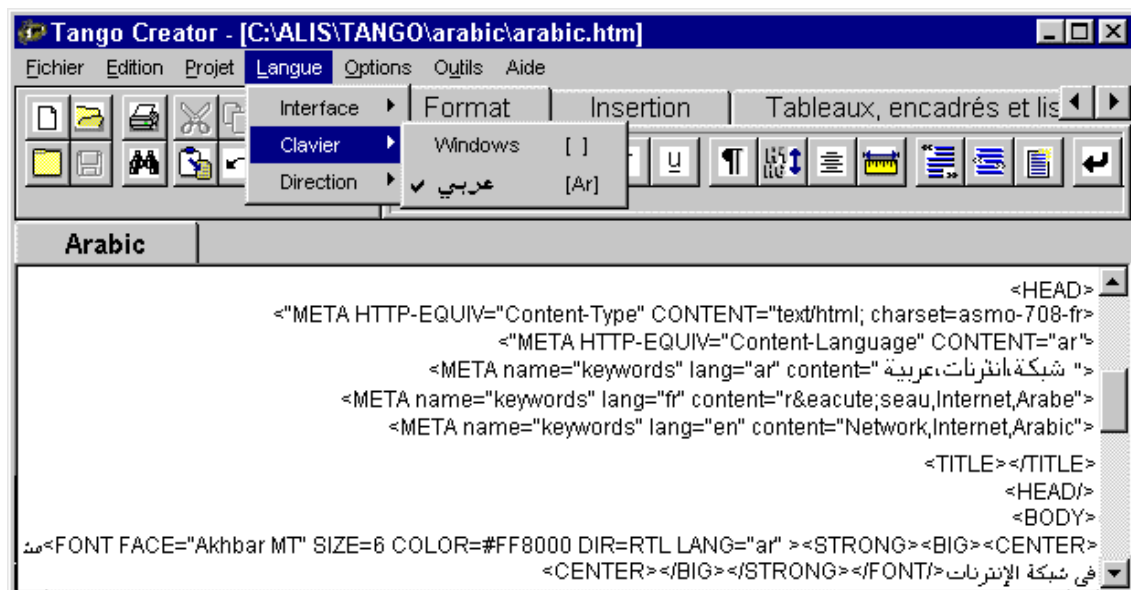


Figure 101 : Indexation et désambiguisation linguistiques par attributs « META name » et « LANG » sous Tango Creator

L'efficacité des moteurs de recherche dans la gestion des versions multilingues des documents peut être améliorée à travers l'usage de l'élément « [LINK](#) » pour spécifier un ou des liens vers des traductions du document dans d'autres langues ou vers des versions différentes du document sur d'autres formes de médias.

4.2.3.3.2 L'attribut « DIR » et le traitement BIDI des documents HTML

La directionnalité de l'écriture est l'un des points forts du multilinguisme que nous avons identifié de "[multilinguisme lourd](#)" dans la problématique de notre présent travail.

Étant donnée que beaucoup de langues sont écrites dans un sens horizontal de gauche à droite et de droite à gauche, la présence simultanée de chaînes de caractères dans deux langues ou plus à directionnalité opposées et dans le même document crée un contexte de bidirectionnalité mieux connu sous l'acronyme BIDI.

Les textes BIDI nécessitent un étiquetage particulier chaque fois qu'une ambiguïté relative à l'orientation du texte écrit est à résoudre. En effet, la difficulté majeure dans la bidirectionnalité

des textes multilingues n'est pas la désignation du sens de l'écriture dans une langue ou dans une autre. Elle réside plutôt dans la concordance de traitement de la bidirectionnalité des textes imbriqués, des directionnalités héritées et des mécanismes qui les annulent. Déjà, les auteurs sont souvent appelés, en plus de la détermination de la langue du document par l'attribut <LANG>, de spécifier le sens de l'orientation de quelques éléments de textes imbriqués ou du texte entier. Cet étiquetage (ou balisage) devrait normalement intervenir pour rendre lisible et compréhensible l'enchevêtrement de textes à orientations diamétralement opposées dans le même document, dans le même paragraphe voire même sur la même ligne. Nous avons déjà abordé le principe de la directionnalité implicite par l'algorithme BIDI sous Unicode. Nous aborderons divers aspects des rendus visuels des textes BIDI et imbriqués sur des interfaces Homme-Machine dans le chapitre suivant. Mais nous verrons dans les paragraphes qui suivent le traitement BIDI par éléments et attributs HTML.

Bien qu'Unicode gère la bidirectionnalité par les caractères de direction (§ III-3.4.2.2), HTML fournit des constructions bidirectionnelles d'un niveau supérieur donnant les mêmes effets BIDI. Parmi les étiquettes que l'Internet a permis d'ajouter au langage HTML, l'élément <BDO> et l'attribut <DIR> (à ne pas confondre avec l'élément <DIR>!) constituent un acquis important pour les auteurs et les utilisateurs afin de résoudre, au niveau de l'interface Homme-Machine, les situations de conflit du sens de l'écriture des textes multilingues et le traitement des incises. Il est ainsi plus convivial de spécifier une directionnalité Droite-Gauche d'un texte arabe par la méthode HTML sous forme :

<Q lang="ar" dir="rtl">...un texte arabe...</Q>

que de faire usage des équivalents Unicode (Tab.27) sous la forme :

‫״...un texte arabe...״‬

L'attribut <DIR> spécifie entre autres les limites d'encapsulation des caractères à directionnalité neutre (espace) ou à valeur directionnelle faible (ponctuation). Les deux valeurs directionnelles de l'attribut <DIR> sont "LTR" pour "Left to Right = Gauche à Droite" ou "RTL" pour "Right To Left = Droite à Gauche".

Pour fixer la direction de base d'un texte dans un document entier, l'attribut <DIR> est défini en conjonction avec l'élément <HTML> comme le démontre l'exemple suivant :

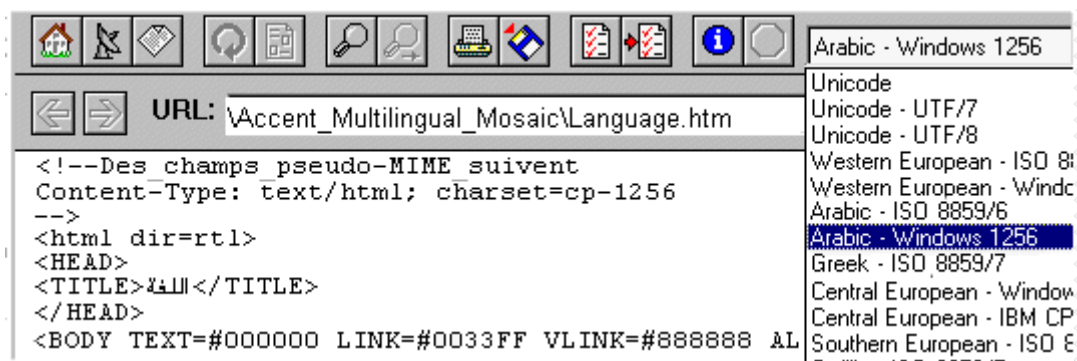


Figure 102 : La direction de base définie par l'attribut <DIR> en jonction avec l'élément <HTML>

Au niveau des blocs de texte, il est toutefois possible de rompre la directionnalité de base du document pour asseoir une directionnalité opposée par le même attribut <DIR>. Chaque fois

qu'elle est omise, cette directionnalité applique le principe de l'héritage directionnel dicté par la directionnalité de l'élément parent précédent. La directionnalité par défaut, en l'absence de toute valeur <DIR> reste toutefois une valeur "LTR" de gauche à droite.

```
<HTML dir="RTL">
...texte droite-gauche...
<P dir="ltr">...texte gauche-droite...</P>
<P>...texte droite-gauche...</P> ←
</HTML>
```

(Directionnalité automatiquement remise à la valeur de l'attribut "DIR" hiérarchique associé à l'élément <HTML>)

Pour clarifier davantage les particularités de l'attribut <DIR>, nous avons conduit la série des tests suivante afin de voir l'effet de certaines combinaisons de traitement BIDI par balises HTML sous différents agents utilisateur. Les tests nous ont également permis de passer en revue certains de ces agents utilisateur et de sonder la profondeur de leur maîtrise du contexte BIDI.

a. TEST 1 : Spécifier la directionnalité de base d'un document HTML

Contexte :

Nous avons soumis le texte échantillon suivant à trois navigateurs Web différents dont deux (02) multilingues (Tango d'Alis Technologies et Navigate with Accent d'Accent Soft) et un (01) latin (MSI-Explorer 4.0). Le texte comprend des paragraphes en latin et en arabe disposés dans une suite intermittente. Une directionnalité Droit-Gauche de base est spécifiée au niveau de l'élément <HTML>.

```
<!doctype html public "-//IETF//DTD HTML 3.2//EN">
<html dir="rtl">
<head>
<title>Document Title</title>
<meta name="Generator" content="HTMled32 Version 2.0a">
<meta name="Author" content="Unknown">
</head>
<body bgcolor="#FFFFFF">
Texte Latin <p>
...ãÕ ÚÑÈí<p>
rédigé en<p>
<b>ÈÇÁÎØ ÇÁÚáíÖ<p>
langue française<p>
</body>
</html>
```

Résultats :

Les plus importantes constatations effectuées sur les rendus des différents produits concernent les deux aspects des fontes et de la directionnalité du texte affiché. Si nous avons déjà évoqué auparavant le problème des caractères, des fontes et des glyphes (voir [Analyse contextuelle](#)) qui nécessiteraient des plates-formes munies des fontes appropriées et d'algorithmes d'analyse contextuelle adéquats (voir [Analyse contextuelle arabe](#)), la bidirectionnalité sous HTML n'est pas moins importante pour des rendus compréhensibles et conformes aux choix linguistiques de l'utilisateur.

Le premier produit testé, en l'occurrence Tango d'Alis, gère convenablement les deux problèmes évoqués et effectue un rendu conforme aux codes sources HTML y compris la valeur "RTL" de l'élément <HTML>.

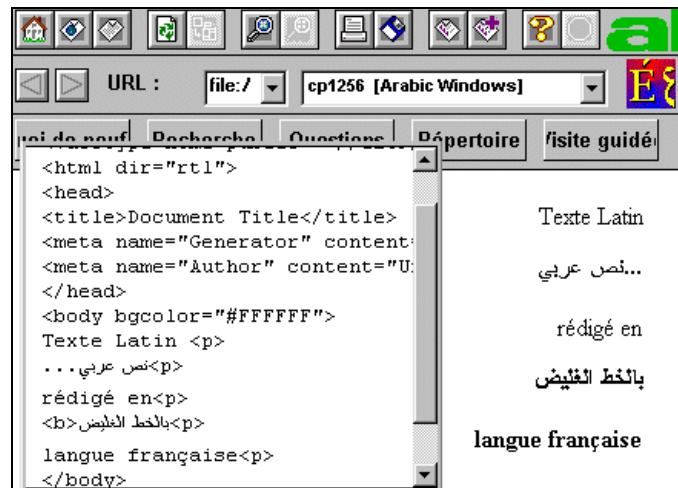


Figure 103 : Conformité de Tango à la valeur "rtl" de la directionnalité de base

Par contre, le navigateur Accent Multilingual Mosaïc, s'il résout le problème des glyphes grâce à la fonte arabe qu'il intègre par défaut, n'a pas été conforme à la directionnalité Droite-Gauche spécifiée par la valeur "RTL" au niveau de l'élément <HTML>. Son rendu est entièrement justifié à gauche.

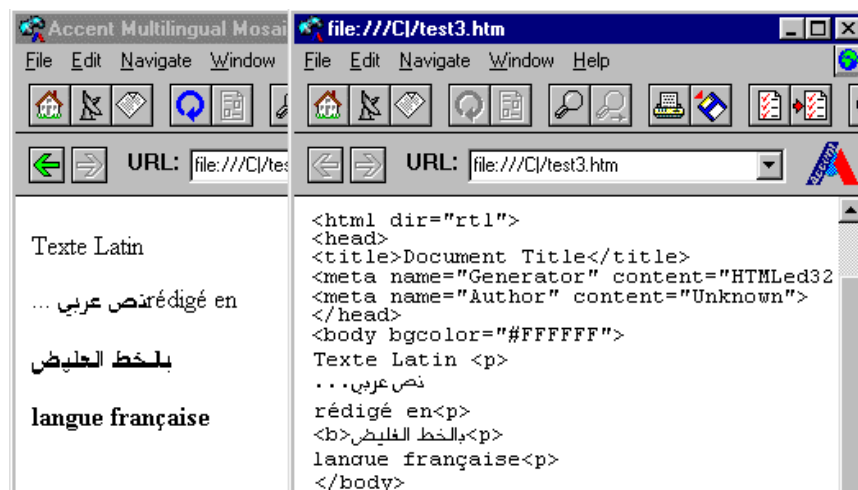


Figure 104 : Inadaptation du navigateur Accent Multilingual Mosaïc au paramètre "RTL"

MSExplorer 4.0, partant de sa nature latine, n'était pas prévu gérer convenablement les deux problèmes de la forme des glyphes ni de la bidirectionnalité du texte.

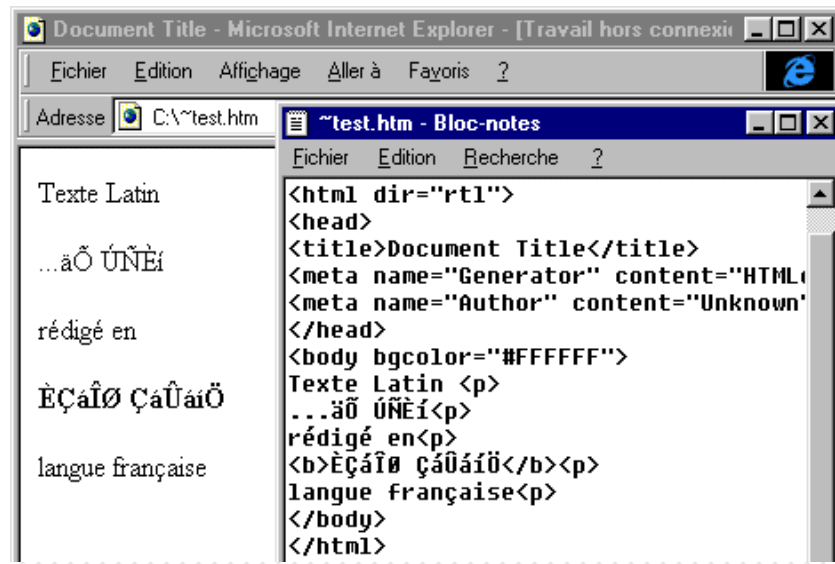


Figure 105 : Double inadéquation de MSIE Explorer Latin 4.0 avec la fonte arabe et la directionnalité de base "RTL"

Des alternatives forcées sont cependant appliquées sur le texte source pour se conformer aux deux spécificités indiquées. D'une part nous avons opté pour des éléments HTML d'une portée non linguistique mais plutôt d'alignement textuel, en l'occurrence l'attribut `<align>`, et d'autre part, nous avons appliqué l'attribut `<font face>` pour le rendu des caractères arabes.

Dans la figure suivante, la valeur "RTL" de l'attribut `<DIR>` n'a eu aucun effet sur un texte latin affiché à gauche. Le rendu du texte arabe, non affecté par la même valeur (Fig.105), a été par contre forcé à un alignement droite-gauche par l'attribut `<align="right">` et à un rendu dans une fonte arabe par l'attribut `<font face="Baghdad">`^[161]

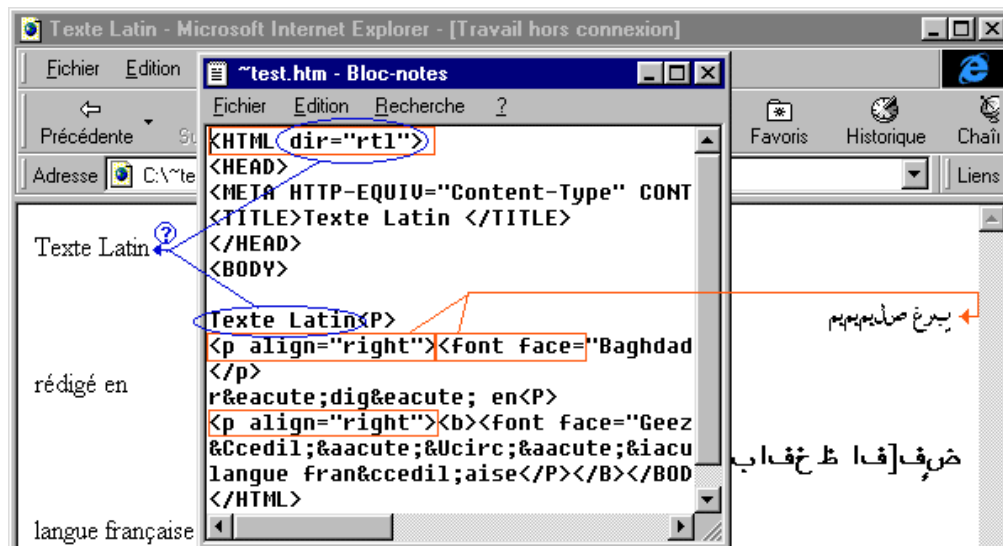


Figure 106 : Alternative "RTL" par attribut `<align="right">` sous MSIE Explorer 4.0

[161] Nous n'avons pas cherché à choisir la fonte correspondante au texte d'origine dans cet exercice. L'essentiel pour nous était de parvenir à un affichage des glyphes arabes.

Cet exemple souligne l'importance et la nécessité de traiter la directionnalité des documents HTML au niveau des blocs de texte pour faire correspondre chaque chaîne de caractères à la nature de son affiliation linguistique.

b. TEST 2 : Spécifier la directionnalité par blocs de texte HTML

Contexte :

L'orientation du texte dans les documents HTML ne peut pas se limiter à la directionnalité de base présentée dans les exemples précédents. La bidirectionnalité nécessite un traitement par blocs de textes conformément à l'orientation de chacune des langues impliquées. Nous avons procédé, par conséquent, à déterminer l'orientation de chaque bloc de texte dans l'échantillon précédent par une déclaration `<dir>` qui conviendrait à sa valeur linguistique.

Résultats :

Le rendu effectué par Tango d'Alis reproduit intégralement la structure interne du document HTML conformément aux spécificités directionnelles de l'attribut `<dir>`. Nous observons dans cet exemple l'application du principe de *l'héritage directionnel* par les blocs de texte latins qui, sans avoir à subir un paramétrage directionnel, héritent de l'orientation gauche-droite implicite de l'élément `<HTML>` situé à un niveau supérieur. Ainsi, les blocs de texte arabes, nécessiteraient un paramétrage explicite par l'attribut `<dir="rtl">` afin d'interrompre l'effet d'orientation de l'élément parent `<HTML>`.

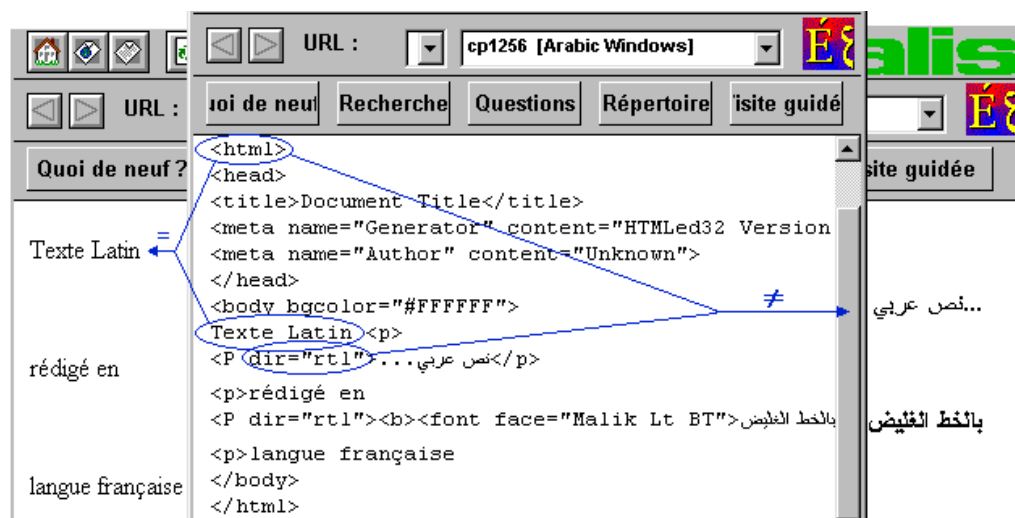


Figure 107 : Conformité intégrale de Tango aux attributs de bidirectionnalité implicites et explicites

Les même variables ont été appliquées aux mêmes blocs de texte sous Accent Multilingual Mosaïc. Le rendu, comme démontré dans la figure suivante, n'a pas été conforme aux spécificités des paramètres HTML de bidirectionnalité. La valeur implicite "LTR" de l'élément `<HTML>` parent est généralisée sur tout le document.

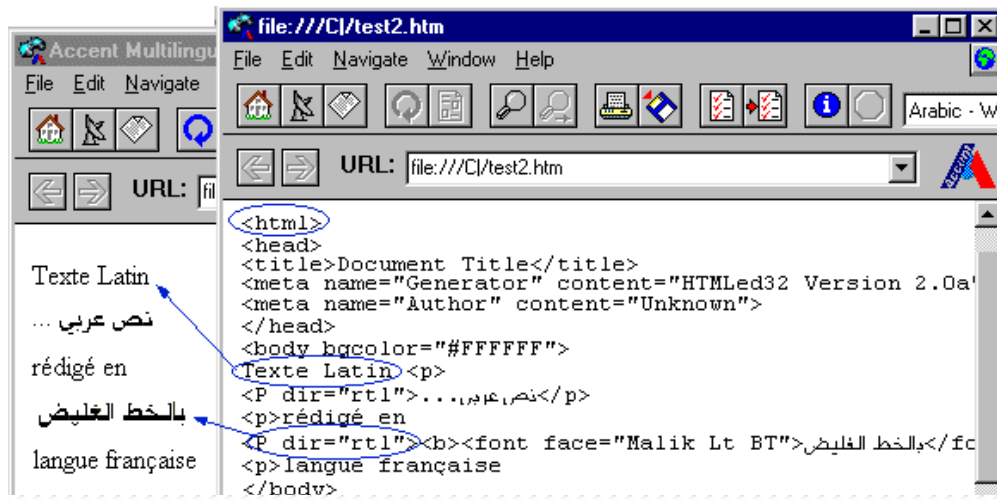


Figure 108 : Non acceptation par Accent Multilingual Mosaic des attributs de directionnalité explicites

Ceci constitue en quelque sorte une défaillance d'Accent Multilingual Mosaic dans l'application de l'algorithme de traitement bidirectionnel. Un balisage explicite de directionnalité par l'attribut `<align>` serait donc nécessaire. La figure suivante illustre le réajustement du rendu bidirectionnel des blocs de texte multilingues sous Accent Multilingual Mosaic en utilisant un balisage d'alignement textuel par l'attribut "align".

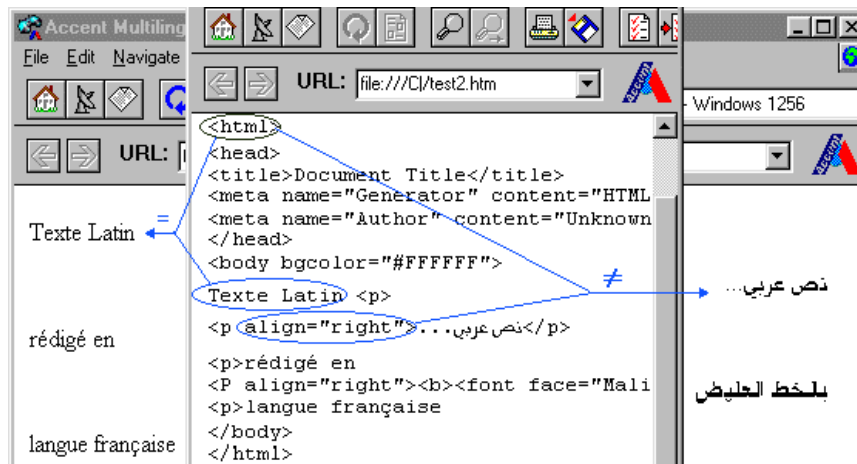


Figure 109 : Réajustement de la bidirectionnalité sous Accent Multilingual Mosaic par attributs d'alignement

c. TEST 3 : Spécifier la directionnalité des incises par niveau d'imbrication hiérarchique

Contexte :

Un niveau de complexité plus profond est quand même méritoire d'une attention plus particulière. L'insertion entrelacée (nested embedding) de caractères à directionnalités opposées, constitue l'une des difficultés majeures du traitement des imbrications et des incises des textes multilingues. C'est l'un des cas où la directionnalité implicite des caractères est insuffisante et donc nécessite un balisage particulier. Nous aborderons ces deux caractéristiques avec plus de détails dans le chapitre suivant (§ IV-2.3.2). Nous nous limiterons dans cette partie aux tournures HTML dans le traitement de cette complexité directionnelle des textes imbriqués.

Nous avons opté pour le navigateur Tango d'Alis comme plate-forme de test étant donné sa conformité optimale démontrée dans les tests précédents avec les pré-requis de la bidirectionnalité.

Nous avons commencé par un cas d'insertion de texte assez simple. Des chaînes de caractères arabes et latins entrelacées dans les deux modes maître d'édition arabe et latin changent d'orientation en fonction de la directionnalité de base implicite ou explicite.

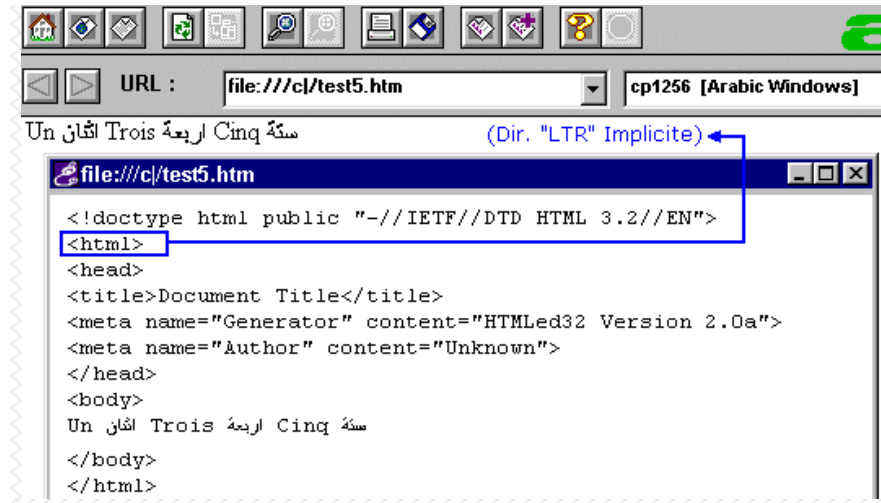


Figure 110 : Directionnalité "LTR" implicite des incises

Dans ce premier cas de figure, l'élément <HTML> le plus générique du document est silencieux sur la directionnalité de base. Tous les éléments subséquents héritent d'une directionnalité implicite de Gauche à droite. L'alignement des chaînes de caractères multilingues s'effectue donc de gauche à droite.

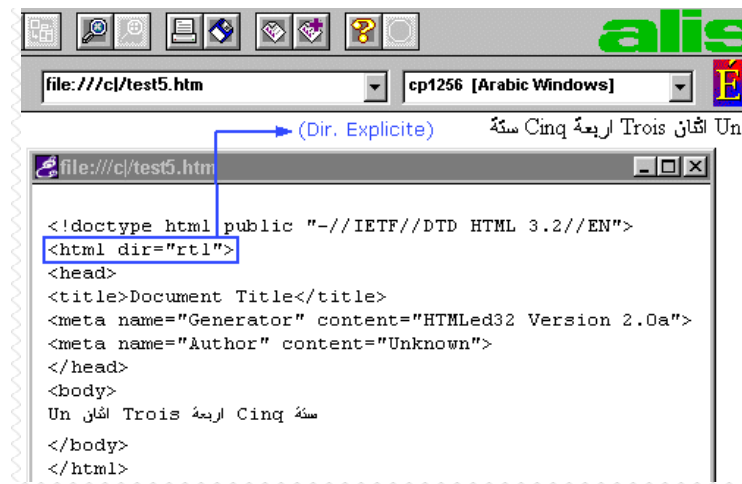


Figure 111 : Directionnalité de base "RTL" explicite des incises

Le passage à un mode Maître "RTL" nécessite une déclaration explicite au niveau de l'élément le plus hiérarchique <HTML>. Tous les éléments subséquents en héritent la directionnalité "RTL".

Jusqu'à ce niveau de complexité des textes BIDI, l'algorithme bidirectionnel d'Unicode qu'utilise HTML et l'attribut <DIR> répondent convenablement aux besoins de traitement bidirectionnel. L'algorithme bidirectionnel Unicode de HTML inverse automatiquement les caractères insérés

en fonction de leur directionnalité implicite ou héritée. Seulement, cette démarche de l'algorithme BIDI ne compte généralement que pour un seul niveau d'insertion. Pour atteindre des niveaux supplémentaires de changement de direction, l'usage de l'attribut explicite <DIR> reste indispensable.

L'exemple suivant est un cas où l'ajout des attributs de direction sur des incises est nécessaire.

```
<SPAN DIR=LTR>Texte Latin<SPAN DIR=RTL>نص عربي<SPAN DIR=LTR>rédigé  
en</SPAN>بالخط الغليظ</SPAN>langue française</SPAN>
```

L'étiquetage par l'attribut étend la zone d'inversion de direction (pivotement=swapping) à toute la zone entre les deux balises d'ouverture et de fermeture de l'attribut . Plusieurs niveaux d'imbrication de zones de texte sont possibles. Nous obtenons ainsi le rendu suivant :

```
Texte Latin الغليظ بالخط rédigé en نص عربي langue française
```

Leur omission résulterait en un rendu différent car la directionnalité "LTR" implicite aurait agité normalement avec l'aide de l'algorithme BIDI pour donner un maximum de deux niveaux de traitement bidirectionnel : celui de la directionnalité de base implicite et une directionnalité à contre sens de l'algorithme BIDI. Ceci donnerait le résultat suivant :

```
Texte Latin نص عربي rédigé en الغليظ بالخط langue française
```

La suppression de la bidirectionnalité (Directional override) est également nécessaire pour contre carter certains comportements non courants de pivotement automatique inapproprié. L'élément <BDO> est utilisé dans des cas de figure où la directionnalité ne pourrait pas être résolue par simple analyse de contexte.

Un exemple concret de cette situation est la valeur numérique composée de chiffres (direction LTR), de points (direction neutre) et de caractères arabes (direction RTL). Pour éviter le pivotement de séquences de caractères qui s'alignent selon l'algorithme BIDI en action, l'attribut <BDO> (BiDirectional Override) force un alignement directionnel continu conforme à la valeur de l'attribut <DIR> dépendant sans aucune considération des propriétés directionnelles des caractères. Il est en quelque sorte équivalent aux caractères LRO (202D) ou RLO (202E) de ISO/IEC 10646 (§ III-3.4.2.2 [RLO/RLE](#)).

Une autre contrainte d'usage de l'attribut <BDO> est le conflit de directionnalité qui pourrait résulter d'un usage simultané avec les codes de formatage BIDI de ISO/IEC 10646. La norme HTML 4.0 recommande l'usage exclusif de l'un ou de l'autre. F. Yergau recommande la technique des balises qui demeure beaucoup plus conviviale pour sauvegarder l'intégrité des structures des documents et permettre de résoudre certains conflits de bidirectionnalité par simple usage d'un éditeur de texte^[162]. L'attribut <BDO> ne devrait être donc utilisé que dans des situations où un contrôle absolu sur la directionnalité du texte est nécessaire (i.e. chiffres multilingues composés).

[162] F. YERGAU. Internationalization of the Hypertext Markup Language. <http://www.sil.org/sqml/rfc2070.html>

5. CONCLUSION

Le contexte de codage de l'information est parsemé, comme démontré dans ce chapitre par une grande multitude de normes et de standards qui s'annulent, se complètent ou se relaient l'un par rapport à l'autre. Ce qui est toujours constant c'est que toutes les démarches normatives naissent d'un consensus autour d'une pratique existante au sein d'une communauté linguistique ou géographique quelconque. Ceci implique à priori un travail de concertation et de coordination entre tous les partenaires concernés afin d'assurer des retombées économiques et informationnelles évidentes. Ceci supposera aussi la présence d'un esprit collectif du besoin de normalisation chez des individus (des personnes ressources) et des institutions qui œuvrent dans des domaines de spécialités diverses. Selon Sylvie Baste, *"l'interaction entre plusieurs domaines entre en jeu dans la construction d'une norme. Le secteur recherche et développement est influencé par le secteur industriel qui est influencé par le secteur marketing, le secteur politico-économique"*^[163]. Elle donne l'exemple du secteur de recherche et développement (R&D) qui subit l'influence des secteurs industriels à leur tour influencés par le secteur marketing.

En terme de normalisation des jeux de caractères sur plusieurs octets, les experts de différentes disciplines ont ainsi un rôle déterminant à jouer dans la démarche normative qui concerne la norme 10646. Avec les deux phénomènes de la localisation et de l'internationalisation des systèmes d'information, la concertation doit enregistrer la contribution de tous les partenaires nationaux et internationaux dans une démarche unifiée et homogène. La prolifération des systèmes ouverts et distribués a aboli les barrières géographiques de manière à rendre interoperables tous les systèmes d'information hétérogènes dans un macrocontexte sans limites dans lequel l'opérande linguistique est censée devenir transparente. Internet et les autoroutes de l'information en constituent l'argument définitif.

À l'heure actuelle, aucun projet de cette envergure universelle n'est réellement existant. Unicode, ISO/IEC 10646 le sont en réalité mais il leur manque encore un grand pas vers l'adoption unanime. Beaucoup de revendications, de propositions et de critiques sont formulées sans qu'une attitude claire et définitive ne soit encore atteinte. Il est évident que toute nouvelle norme passe par une phase de test (le drafting ou bien les RFC) et que les premières versions ne sont pas conformes aux attentes de tous les partenaires. Le duo Unicode-ISO/IEC 10646 cherche encore les voies de l'adoption totale et universelle. Les différentes mises à jour et l'édition des errata des normes en question dénotent d'une volonté de porter ces alternatives au degré de l'adoption totale et uniforme.

Dans l'attente d'arriver à ce stade d'universalité pratique des normes de codification à 16 et 32 bits, il est encore temps de traiter davantage les normes dites "historiques" à 7 et 8 bits et de développer les tables de leur conversion vers et de l'Unicode. En se positionnant en amont des normes 16 et 32 bits, des solutions d'adaptabilité s'appliquent déjà. Le renforcement du passage du codage des caractères de 8 bits à celui de 16 bits constitue aujourd'hui une première démarche de la transition intégrale vers le codage à 32 bits. Une première étape de codification sur 16 bits permettrait en effet de préparer le terrain pour un codage unique à 32 bits. Car il serait, selon certaines opinions, encore prématuré de se lancer vers une codification canonique à 32 bits avant d'atteindre un consensus général et une envergure universelle basée sur l'équité dans la distribution des espaces de codes et l'exhaustivité dans la représentation linguistique de toutes les langues du monde (écrites et orales).

[163] Sylvie Baste. Recherche sur la normalisation internationale du codage des caractères codés sur plusieurs octets. Novembre 1997.

CHAPITRE QUATRIEME

CATALYSEURS ERGONOMIQUES DES INTERFACES HOMME-MACHINE DANS LES SYSTEMES D'INFORMATION MULTILINGUES ARABE/LATIN

Il nous semble nécessaire à ce stade d'avancement de cette étude, de cerner davantage notre sujet en amont et en aval de ses points de culmination ; en l'occurrence les interfaces bilingues et leur rôle dans la communicabilité de l'information scientifique et technique. Ceci nous met d'emblée devant une nécessité de procéder à un approfondissement dans la structure scripto-graphique de la langue arabe comme protagoniste à part entière qui accompagne déjà les langues latines au standard des IHMs bilingues. La mise en exergue de certains aspects de cette structure scripto-graphique saurait expliquer, tout au long de ce travail, certains aspects de rigueur ou de souplesse de la langue arabe dans son côtoiement aux langues latines.

Malgré la dissolution de l'ASMO en 1990 qui justifierait en quelque sorte l'éclatement de l'effort de normalisation arabe et sa soumission aux aléas des tentatives éparses nationales et étrangères, une première critique pourrait à notre sens être formulée à l'égard des versions définitives de l'ASMO 449 et 708 ainsi que leurs équivalentes internationales de l'ISO en ce qui concerne l'allocation des codes de caractères et la définition du jeu définitif des éléments des codes graphiques.

Le caractère arabe en fait connaît des particularités graphiques qui l'identifie comme une langue orientale cursive qui s'écrit de droite à gauche (DAG=RTL). Cette caractéristique cursive constitue, selon Ghzal^[1], l'une des formes de son état archaïque qui, à l'opposé de l'écriture latine, ne s'est pas encore mise au diapason de l'industrialisation pour s'adapter à l'outil technologique. Un texte latin aurait évidemment deux formes distinctes selon la méthode de son écriture : manuscrite cursive (rattachée) ou imprimé isolé. Les algorithmes de l'analyse contextuelle n'y sont appliqués que pour certaines situations relatives aux ligatures et aux caractères composés (æ, Æ, œ, ß...) ou dans des situations plus avancées pour la correction orthographique et la vérification grammaticale. D'ailleurs, toute une polémique concernant l'adaptation de l'outil à la langue ou l'adaptation de la langue à l'outil a constitué un débat culturel très profond entre les techniciens et les linguistes dans le monde arabe. Des "pours" et des "contres" ont marqué ce discours pour prêcher d'une part la dynamisation du caractère arabe et son introduction dans les sphères technologiques et informationnelles mondiales et d'autre part, pour protéger l'originalité de la langue arabe et sa richesse calligraphique et artistique contre les déformations de l'écriture électronique. Une première distinction fondamentale de concepts, souvent source de confusion, concernant toute langue écrite est d'ores et déjà indispensable : la nuance entre les caractères et les glyphes.

1. LA COMPOSITION DU TEXTE ÉCRIT : ENTRE LE SENS DU CARACTERE ET LA VISIBILITÉ DU GLYPHE

Tout processus d'affichage ou d'impression est en fait un processus de rendu^[2] (visuel ou graphique) qui transpose une suite de caractères en une suite de glyphes rendus visibles par les

[1] Ahmed EL AKHDAR GHOZAL. Les appareils d'impression électronique. L'ordinateur et le caractère arabe : problématique et solution. pp.244-354. (En langue arabe)

[2] Le terme rendu est utilisé par Unicode pour désigner les opérations d'affichage ou d'impression des caractères. Nous en ferons usage pour désigner les mêmes opérations sous d'autre normes de codage.

équipements de traitement en place. « Les glyphes sont les représentations visuelles des caractères qui sont contenues dans les polices ou fontes, alors que les caractères sont les éléments logiques de base du texte »^[3]. C'est là une définition qui pose la problématique véritable issue de cette distinction de base.

Caractères et glyphes sont en réalité deux concepts qui prêtent à confusion bien qu'ils partagent plusieurs attributs en commun. Leur distinction en tant qu'entités différentes devient essentielle dans le traitement automatisé de l'information.

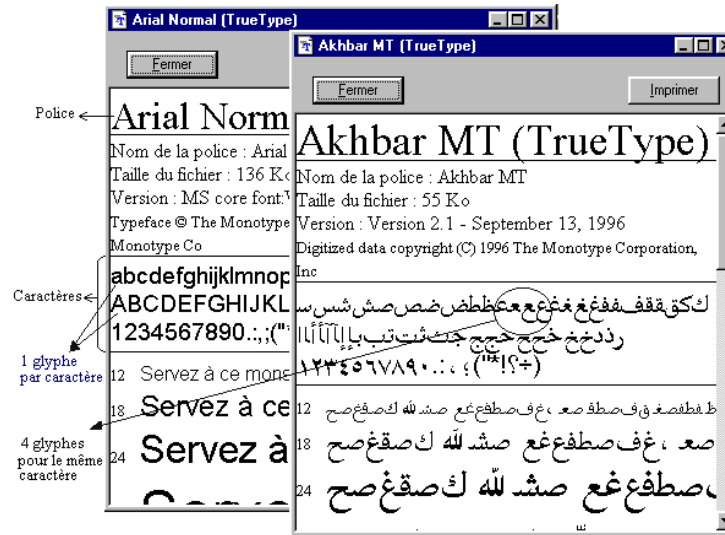


Figure 112 : Les nuances Police/caractère/glyphe

La distinction entre caractères, polices de caractères, fontes et glyphes émane en réalité du milieu de l'imprimerie. Les instances de normalisation, comme l'ISO les ont repris dans leurs textes de normes. Selon la nomenclature de l'ISO/IEC 10646 et de l'ISO/IEC 9541, les définitions suivantes s'appliqueront au contexte de notre étude.

- ❑ Un caractère fait partie d'un jeu d'éléments utilisés pour l'organisation, le contrôle ou la représentation des données. (ISO/IEC 10646-1:1993)
- ❑ Un jeu de caractères codés est un ensemble de règles non ambiguës qui définissent un jeu de caractères ainsi que les relations qui les unissent et les codes qui les représentent. (ISO/IEC 10646-1:1993)
- ❑ Une fonte est une collection de glyphes ayant la même forme calligraphique (i.e. Times New Roman, Gras, Italique ...). (ISO/IEC 9541-1:1991) C'est un jeu de caractères d'une police de caractères. Les deux termes sont souvent utilisés comme synonymes. (Fig.66)
- ❑ Un glyphe est une forme abstraite d'un symbole graphique. (ISO/IEC 9541-1:1991)
- ❑ Un caractère graphique, à l'encontre des caractères de contrôle, dispose d'une représentation visuelle manuscrite, imprimée ou affichée. (ISO/IEC 10646-1:1993)
- ❑ Un symbole graphique est la représentation visuelle d'un caractère graphique ou d'une combinaison de caractères graphiques. C'est l'équivalent du glyphe. (ISO/IEC 10646-1:1993)

L'ISO/IEC 10646 établit une distinction entre les deux concepts par la désignation de la représentation visuelle des caractères comme "symbole graphique". Symboles graphiques et glyphes sont des synonymes avec la distinction que ce dernier est adopté dans la nomenclature

[3] <http://www.alis.com>

de la norme ISO/IEC 9541^[4]. Une meilleure définition de cette distinction entre caractère et glyphe et la nature de la relation qui les unit peut être formulée comme suit :

Un caractère implique une distinction de sens ou de son. Entre une "A" et une "B" la signification et la prononciation changent. Par contre entre « A » et « a » ni le sens du caractère ni sa valeur phonétique ne changent. Selon cette distinction faudrait-il donc considérer « A » et « a » comme caractères ou glyphes surtout que chacun dispose de son propre point de code dans les tables de codification ? Sachant que les tables de code attribuent des éléments de codes à des caractères et des glyphes à la fois, l'application de la définition précédente en fait deux glyphes d'un même caractère. Ainsi le caractère ne dispose d'aucune forme visuelle intrinsèque.

Un glyphe implique une distinction de forme. Si la lettre « A » et la lettre « a » sont deux glyphes d'un même sens et d'une même prononciation d'un même caractère, ils ne disposent cependant pas de la même forme visuelle. Pourtant ils disposent de deux points de codes différents. L'écriture arabe, étant son aspect cursif, est entièrement basée sur la notion de glyphes. La majorité des caractères des tables des codes sont traités au niveau du rendu graphique par l'analyse contextuelle, alors que certains autres (le cas des 5 formes de la « Hamza » positionnés de 4/2 à 4/7 dans l'ASMO 449 et de 12/02 à 12/07 dans l'ASMO 708), sont en réalité des glyphes d'un même caractères.

Un ou plusieurs caractères peuvent être représentés par un ou plusieurs glyphes. ISO/IEC 10646 définit dans ce sens 4 possibilités de rapports entre caractères et glyphes :

Un à un : C'est le mode de représentation le plus utilisé pour les langues latines. Le caractère UCS U+0041 équivalent à la lettre majuscule « A » est représentée par le glyphe « A ». Le caractère « a » est également représenté par le glyphe « a ». Il est toujours possible qu'un seul glyphe représente plus d'un seul caractère. Le glyphe Å (« A » capitale avec un anneau dessus) représente à la fois le caractère des langues nordiques et le signe Angström^[5]. Ce cas est à nuancer du cas « Plusieurs-à-un » traité ci-après.

La liaison un-à-un n'est pas appliquée en langue arabe car chaque caractère dispose d'une variété de glyphes définis en fonction de sa position dans le mot. Dans le tableau 34 ([Tab.34](#)), chaque caractère arabe dispose de plus d'une forme visuelle choisie par l'algorithme de l'analyse contextuelle.

Plusieurs à un : C'est aussi un modèle très utilisé dans les langues latines surtout au niveau des caractères combinés comme les diacritiques. Le glyphe « é » dans la norme ISO/IEC 10646 est la combinaison des deux codes des caractères U+0065 « e : lettre e minuscule » et U+0301 « ' : accent aigu ». Le cas du glyphe ½ constitué des trois caractères 1, / et 2 en est un autre exemple. Cette méthode est également très utilisée dans les écritures non latines où l'usage des diacritiques est plus fréquent. Le glyphe arabe « Lam-Alef » est en fait une combinaison des deux codes des caractères « Lam » et « Alef » (Voir la problématique de ce caractère ci-après).

Un à plusieurs : Cette caractéristique de représenter un seul caractère par deux glyphes à la fois est plutôt une particularité des langues asiatiques où les voyelles peuvent être écrites en deux morceaux. La voyelle bengale U+09CB peut être affichée moyennant deux glyphes de part et d'autre de la consonne articulée. L'écriture arabe présente le cas des 5 formes de la « Hamza »

[4] ISO/IEC 9541-1:1991 Information Technology -- Font Information interchange -- Part 1 : Architecture.

[5] Unité de mesure de longueur d'onde et des dimensions atomiques (symb. Å), valant un dix-milliardième de mètre (10⁻¹⁰ m).^[5]

qui constituent, selon notre point de vue, un cas d'anomalie que nous détaillons dans les paragraphes suivants (Voir le point 2.2).

Plusieurs à plusieurs : C'est l'alternative la moins fréquente. Nous citons cependant le cas de l'écriture vietnamienne qui permet la combinaison des caractères UCS U+0065 « e », U+0302 « ^ » et U+0323 « . » afin d'engendrer le glyphe « e » avec un accent circonflexe et un point en dessous. Dans des systèmes qui disposent déjà du caractère « ê », il leur suffit de le combiner avec le code du point pour avoir le même glyphe de l'exemple précédent. Cette situation est également possible en langue arabe avec des caractères composés du genre « Lam-Alef » voyellé ([fig.113](#)). « Lam-Alef » qui ne figure nulle part dans les tables de code ASMO et leurs équivalentes ISO, sauf exception dans la page de fontes 165 sous MS-DOS au niveau des emplacements entre 90/09 et 90/0E (en hexadécimal) et entre F0/09 et F0/0A et la page de code 1256 sous Windows aux emplacements numérotés entre 220 et 227 ([Tab.33](#)), peut être composé, comme illustré dans la figure suivante, selon plusieurs méthodes à partir des codes des caractères individuels.

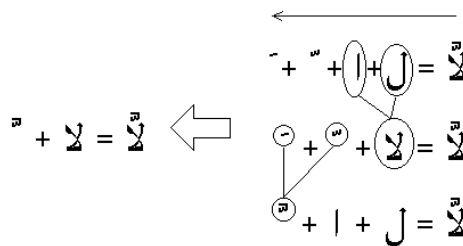


Figure 113 : La correspondance plusieurs à plusieurs entre caractères et glyphes : cas du "Lam-Alef"

Ce ne sont là que quelques-unes des caractéristiques de distinction entre les deux concepts de caractère et de glyphe. Il est en effet possible d'observer en profondeur cette distinction au niveau de certains phénomènes complexes du processus du rendu (visuel ou imprimé) comme les formes contextuelles d'un seul caractère en fonction de son emplacement dans le mot, la ligature de certains caractères, l'inversion, le réagencement bidirectionnel de l'écriture, l'adjonction ou l'omission des traitements des diacritiques. Dans tous ces états de complexité dans le traitement des langues, il ne s'agit plus d'un simple choix du glyphe correspondant au caractère codé dans une police bien déterminée, mais plutôt d'une analyse de tout le contexte environnant où caractères et glyphes sont censés prendre place. L'intervention d'algorithmes appropriés est nécessaire à ce stade pour résoudre des problématiques inhérentes aux cas de complexités mentionnés.

2. LES COMPLEXITES D'USAGE DE L'ÉCRITURE ARABE

L'écriture arabe présente quatre de ces aspects de complexités du rendu visuel des caractères : la forme contextuelle et le choix des glyphes, le réagencement bidirectionnel et la gestion des incises, les ligatures et les caractères composés, la voyellation et le traitement des diacritiques. Ces particularités ainsi que d'autres relatives aux anomalies de codage, à la dualité entre chiffres arabes et hindous, aux problèmes de tri... feraient l'objet de la partie suivante.

2.1 LES CODES DES CARACTERES ET LES LACUNES DES NORMES DE CODIFICATION

La langue arabe est constituée de 28 lettres (consonnes) standards selon l'ordre linguistique et philologique conventionnel ([Tab.32](#)). Les voyelles arabes, très peu utilisées pour la fonction de

« Tachkil » ou voyellation, sont d'un nombre de 5 voyelles simples (Hamza, Dhamma, Kasra, Soukoun et Shadda) et trois composées qui servent le « Tanouin » ou terminaison nasale (Fathatan, Dhammaton, Kasratin).

Dans un usage courant, les voyelles ne sont utilisées que pour discerner certaines similitudes syntaxiques, grammaticales ou sémantiques qui peuvent engendrer un changement de sens ou d'attribut grammatical parfois basé sur la nature et l'emplacement d'un seul caractère diacritique dans le mot. Pour la majorité des utilisateurs avertis, la confusion sémantique, grammaticale ou syntaxique est généralement résolue à travers le contexte. Dans l'exemple suivant, la phrase peut être lue et analysée par un lecteur arabophone sans besoin de diacritiques. Pour des problèmes d'ambiguïté, une « diacritisation » partielle servirait à dissiper l'ambiguïté sémantique, syntaxique ou grammaticale^[6].

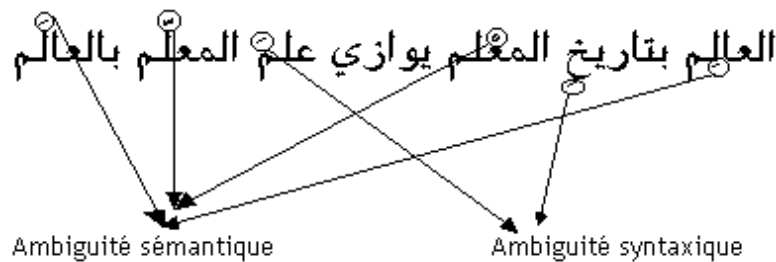


Figure 114 : Désambiguïsation par diacritisation partielle

Parmi les problèmes qu'engendrent les diacritiques partielles, les incohérences de tri et les bruits et silences de la recherche constituent les handicaps majeurs auxquels font face les systèmes d'indexation et de recherche sur texte partiellement « diacritisé ». Pourtant, comme le démontreront les tables des codes de la langue arabe, les voyelles sont prises en compte dans tous systèmes d'information automatisée bien qu'une grande proportion d'utilisateurs n'en fassent aucun usage.

La table ASMO 449 et les tables ASMO étendues signalent en fait 36 caractères différents. Cette différence est en réalité issue de la considération de plusieurs glyphes de certaines lettres comme caractères indépendants. Il s'agit des trois lettres « Hamza », « TA » et « YE » qui prennent des formes variées comme illustré dans le tableau suivant.

Tab. Unicode	Tab. ASMO 708	Tab. ASMO 449	Formes	Nom
0621	12/01	4/1	ء	Hamza
0622	12/02	4/2	آ	
0623	12/03	4/3	إ	
0624	12/04	4/4	ؤ	
0625	12/05	4/5	أ	
0626	12/06	4/6	آ	
0629	12/09	4/09	ة	TA
062A	12/10	4/10	ت	
0649	14/09	6/9	ي	YA
064A	14/10	6/10	ي	

Tableau 31 : Des glyphes / caractères dans les tables des codes arabes

[6] Traduction de la phrase arabe : « Le savoir de celui qui connaît l'histoire du monument équivaut le savoir de l'instituteur pour l'univers ».

La structure de la table ASMO 449 était définie sous inspiration de la table ASCII devenue une référence internationale dans l'assignation des 128 caractères de base dans toutes les tables des codes de l'ISO à 7 ou à 8 bits. La seule différence est au niveau de la représentation des caractères graphiques pour chaque langue dans les zones GG (codage à 7 bits) ou GD (codage à 8 bits). L'objectif dans la distribution des codes de caractères selon la disposition actuelle des tables des codes vise essentiellement la garantie d'un tri binaire des données saisies ou transmises sans nécessité de recourir à des programmes complexes. L'ASMO 449 est donc calquée sur la table ASCII de sorte qu'à la « Hamza », première lettre arabe, est attribué le même code que la lettre « A » dans la table ASCII, à la seconde lettre, en l'occurrence la « BA », est attribuée la valeur binaire de la lettre « B » etc. Le problème est que les lettres de l'alphabet arabe reprises par ASMO 449 dépassent en nombre les lettres de l'alphabet latin signalées dans la table ASCII par 10 positions. La solution était de faire attribuer à cet excédent de 10 lettres les codes des lettres minuscules de la table ASCII. C'est pour cette raison que la lettre « FA » occupe la position de la lettre « a » minuscule et la dernière lettre « YE » occupe la position du code de la lettre « j » minuscule. Ainsi l'ordre alphabétique des caractères arabes s'est trouvé interrompu par six emplacements attribués aux caractères spéciaux de 5/11 à 6/0 dans les tables à 7 bits et 13/11 à 14/00 dans les tables à 8 bits. Ceci est d'ordre à perturber l'ordre de tri qui se trouve ainsi contraint de recourir à des algorithmes spécialisés.

Nom	Lettre	N°	Nom	Lettre	N°
dha	ض	15	alif	ا	1
'ta	ط	16	ba	ب	2
'tha	ظ	17	ta	ت	3
'ein	ع	18	tha	ث	4
ghain	غ	19	jiim	ج	5
fa	ف	20	'ha	ح	6
qaf	ق	21	kha	خ	7
kaf	ك	22	dal	د	8
lam	ل	23	thal	ذ	9
miim	م	24	ra	ر	10
noun	ن	25	zain	ز	11
ha	هـ	26	seen	س	12
waw	و	27	sheen	ش	13
ye	ي	28	'sad	ص	14

Tableau 32 : Les 28 lettres de l'alphabet arabe

Avec l'extension des tables des codes arabes à 8 bits, la combinaison des deux tables des codes à 7 bits, à savoir ISO 646 et ASMO 446, a donné lieu à une table commune à 8 bits (ASMO 708 = ISO 8859-6) ou ISO 646 occupe la partie gauche de la table des codes et ASMO 646 occupe la partie droite sans reprise des caractères de contrôles, de la majorité des caractères spéciaux et des 10 chiffres. Ces caractères sont en fait communs aux deux langues donc il était inutile de les reproduire dans la partie droite. Seules exceptions à cette règle, 6 caractères contextuels arabes ont été miroités sur leurs équivalents latins et introduits dans la partie droite de la table pour satisfaire des exigences de contextualité nécessaires pour le traitement de la langue arabe. Il s'agit des caractères suivants:

1. NBSP (Non Back Spacing) = Espace sans coupure (10/00) : *Caractère graphique dont la représentation visuelle consiste en l'absence de symbole graphique, à utiliser lorsqu'une coupure de ligne est à éviter dans le texte présenté* ^[7].

[7] ISO 8859-6. 1 éd. du 15-08-1987.

2. Symbole monétaire international (10/04) : Ce symbole remplace le symbole dollar (\$) et sert pour la désignation des symboles monétaires nationaux.
3. Virgule arabe (10/12) : La virgule arabe est obtenue par un pivotement horizontal puis vertical de la virgule latine. Le pivotement horizontal coïncide avec l'inversion de la directionnalité de l'écriture. Le pivotement horizontal est quant à lui héréditaire de la calligraphie arabe et de la tradition manuscrite de l'écriture arabe qui voit en l'acte graphique une linéarité cursive et une continuité de l'acte de la main dans le tracé des caractères arabes. En observant la structure des glyphes arabes en positions isolés ou en terminaisons de mots, là où la virgule prend normalement place, tous les glyphes (sauf 6 d'entre eux) ont une terminaison de bas en haut. Une virgule en arabe, compte tenu de la largeur de sa base, s'écrit de bas en haut sous forme d'un point étiré en oblique vers le haut à droite.
4. SHY = Tiret temporaire (10/13) : *Caractère graphique représenté par un symbole graphique identique ou semblable à celui qui représente le TIRET, à utiliser lorsqu'une coupure de ligne a été opérée dans un mot.*
5. Point Virgule arabe (11/11) : Les mêmes observations pour la virgule arabe s'appliquent au point virgule arabe.
6. Point d'interrogation arabe (11/15) : Le point d'interrogation arabe est obtenu par le simple pivotement horizontal du point d'interrogation latin.
7. La « Kashida » ou « Tatwil » (14/00) : Caractère utilisé pour allonger l'espace entre deux caractères pour des raisons de justification^[8] ou d'aération de texte^[9].

L'évolution des normes de codage de caractères à abouti à une situation où la norme ISO 8859-x s'est largement répandue à travers le monde entier comme réponse à l'hégémonie de l'ASCII sur tous les niveaux de la conception des programmes, de la codification de l'information ou de la transmission des données. En fonction de cette tendance d'internationalisation, le caractère arabe s'est vu aussi codé unanimement selon la norme ISO/IEC 8859-6 (équivalente à la recommandation ECMA 114 et à la norme ASMO 708). Bien que la majorité des applications Unix, PC et Macintosh aient adopté totalement ou partiellement cette norme internationale, d'autres jeux de caractères codés pour la langue arabe continuent à être utilisés sous forme de "code page" (Voir [Annexes D1, D2](#)) dans des environnements PC ou dans des installations particulières à des constructeurs bien déterminés (IBM, Wang...). Les pages de codes fonctionnels sous MS-DOS CP 709, CP 710, CP 711, CP 720, CP 786, CP 864, CP 868 et tant d'autres constituent aujourd'hui les alternatives en cours de la codification du caractère arabe dans une plate-forme PC.



Figure 115 : Choix de pages de codes multiples sous MS-DOS

[8] Nous verrons dans la partie réservée à la bidirectionnalité que le texte arabe, contrairement au texte latin, est justifié à gauche et à droite par l'insertion de tirets (Kashida) entre les caractères à l'intérieur des mots au lieu d'ajuster les espaces qui les séparent.

[9] Le « Tatwil » ou prolongement est l'un des héritage de l'écriture manuscrite et de la calligraphie arabes qui pratiquent énormément le « Tatwil » pour des raisons d'aération ou d'esthétique.

très limitatives. Des jeux supplémentaires incluant toutes les formes possibles de représentation des caractères sont indispensables. Ces jeux supplémentaires sont appelés les pages de fontes.

Une page de fonte est donc une reproduction de toutes les permutations possibles d'un caractère qui peuvent changer en fonction de son contexte d'occurrence isolé, au début, au centre ou la fin d'une chaîne de caractères. Les pages de codes ne contiennent généralement que la première de ces 4 permutations. Les pages de codes les plus utilisées sous Arabic MS DOS sont les CP 708, 709, 710 et 864 en combinaisons croisées avec les pages de fontes 161, 162, 163, 164 et 165^[10].

Isolé	Finale Jonctive	Milieu	Début		Isolé	Finale Jonctive	Milieu	Début	
غ	غ	غ	غ	21	ء	ء	ء	ء	1
ف	ف	ف	ف	22	ئ	ئ	ئ	ئ	2
ق	ق	ق	ق	23	و	و	و	و	3
ك	ك	ك	ك	24	ب	ب	ب	ب	4
ل	ل	ل	ل	25	ت	ت	ت	ت	5
م	م	م	م	26	ث	ث	ث	ث	6
ن	ن	ن	ن	27	ج	ج	ج	ج	7
ه	ه	ه	ه	28	ح	ح	ح	ح	8
ة	ة	ة	ة	29	خ	خ	خ	خ	9
ى	ى	ى	ى	30	د	د	د	د	10
ا	ا	ا	ا	31	ذ	ذ	ذ	ذ	11
أ	أ	أ	أ	32	ر	ر	ر	ر	12
إ	إ	إ	إ	33	ز	ز	ز	ز	13
آ	آ	آ	آ	34	س	س	س	س	14
لا	لا	لا	لا	35	ش	ش	ش	ش	15
لأ	لأ	لأ	لأ	36	ص	ص	ص	ص	16
لا	لا	لا	لا	37	ض	ض	ض	ض	17
لا	لا	لا	لا	38	ط	ط	ط	ط	18
و	و	و	و	39	ظ	ظ	ظ	ظ	19
ي	ي	ي	ي	40	ع	ع	ع	ع	20

Tableau 34 : Les caractères arabes et leurs formes contextuelles conformes à ASMO 969

Des solutions de conversion (mapping) de et vers les standards internationaux demeurent cependant très fréquentes. La technique des pages de codes est en réalité une forme de diversité déroutante et limitative de codage car 2 ou 4 tables des codes au plus peuvent être permutées à la fois selon les techniques d'extension de la norme ISO 2022. Unicode et ISO/IEC 10646 essayent de remédier à cette limitation par l'instauration d'une nouvelle distinction entre les concepts de caractère et de glyphe et par l'ouverture des nouveaux systèmes sur le répertoire international de caractères. Déjà les premières solutions d'applications multilingues intégrant l'arabe sur la base de l'Unicode sont apparues. Nous en étudierons quelques exemples dans le chapitre suivant réservé à l'évaluation des interfaces multilingues Homme-Machine. Les tables d'équivalence (mapping) entre l'ISO/IEC 8859-6 (ASMO 708) et l'Unicode sont déjà disponibles (Voir [Annexe H](#)) pour assurer la transparence entre les normes de codage en vigueur. Désormais, il n'est plus question que de la volonté des concepteurs de logiciels et de fabricants de matériel pour adhérer à ce nouveau mode d'internationalisation des solutions informatiques pour le traitement des données multilingues.

À l'opposé du standard latin, la forme des glyphes représentés par les tables des codes arabes ou leurs équivalentes internationales ne constitue pas l'identité réelle des caractères arabes. Conformément à la particularité de l'écriture arabe, chaque caractère dispose d'un nombre variable de formes différentes (glyphes) définies en fonction de son emplacement dans le mot ([Tab.33](#)) :

[10] Voir le manuel Microsoft MS DOS User's Guide : Arabic Supplement. 1988 pour plus de détails.

- ❑ Une forme de début : le caractère constitue le commencement du mot ;
- ❑ Une forme du milieu : le caractère est entouré par d'autres caractères arabes dans le même mot ;
- ❑ Une forme de fin : le caractère est situé à la fin du mot ;
- ❑ Une forme isolée : le caractère est entouré de blancs (et non d'espaces) dans la chaîne de caractères qui constituent le mot.

Cette variabilité du rendu visuel ou imprimé du même caractère est en fait définie par la norme ASMO 969 qui détermine les glyphes de début et de fin d'un même caractère dans une table de fontes constituée de 74 formes des 28 lettres de base et des diacritiques qui les accompagnent (« FATHA », « DHAMMA », « KASRA »)^[11], quatre pour le glyphe « Alef-Lam », dix pour les chiffres de 0 à 9 et 27 pour les caractères spéciaux (Tab.35). L'exception faite pour les lettres « AIN » et « GHAIN » dont les 4 formes des glyphes du début, milieu, fin et isolé sont signalée dans la table minimale de la dite norme, est issue du fait que la forme du glyphe au milieu du mot change complètement d'apparence. Les lettres « DA », « DHA », « RA », « ZAD », « TAD », « DHAD », « WAW » et « ZAIN » sont représentés par un seul glyphe qui est identique dans les deux positions au début d'un mot ou isolé. La « HAMZA » est unique dans son genre car elle nécessite un support (un « ALEF », un « WAW » ou un « YE » en fonction de règles strictes expliquées ci-après) pour apparaître dans les quatre positions de début, milieu, fin d'un mot ou isolé. La version optimale d'ASMO 969 consiste à reproduire toutes les permutations possibles des caractères et des glyphes par l'adjonction à la version minimale des glyphes du milieu auxquels est ajouté un symbole de ligature « KASHIDA » sous forme d'un tiret de jonction. Nous y retrouvons également les différentes formes du glyphe « Lam-Alef » que nous étudierons aussi ci-après.

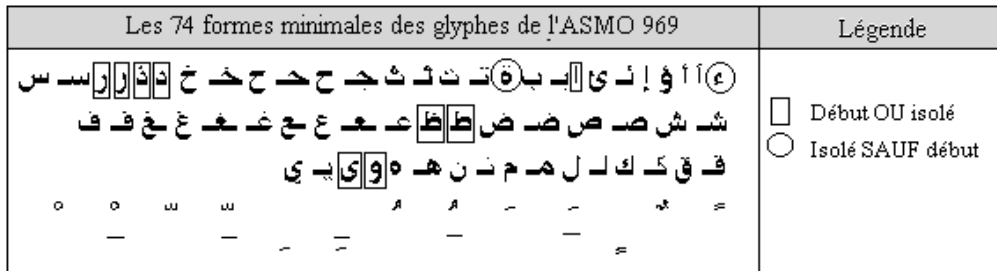
Les 74 formes minimales des glyphes de l'ASMO 969	Légende
	□ Début OU isolé ○ Isolé SAUF début

Tableau 35 : Le jeu minimal des formes de caractères arabes selon ASMO 969

La particularité cursive de l'écriture arabe est la source principale de cette complexité de l'acte scriptural par des procédés électroniques. La forme de l'écriture arabe par les outils mécaniques (machines à écrire) ou électroniques (ordinateurs) est en fait le prolongement inaltéré de la forme de l'écriture arabe manuscrite dans laquelle les caractères sont liés l'un à l'autre selon des règles linguistiques bien déterminées. C'est le principe de la ligature qui présente les trois possibilités suivantes :

- ❑ Certains caractères peuvent accepter une ligature sur les deux cotés. Ils disposent ainsi de 4 possibilités de glyphes.
- ❑ Quelques caractères ne peuvent être liés que sur leur coté droit (2 glyphes possibles),
- ❑ Certains autres caractères ne peuvent accepter aucune forme de ligature. Ils restent ainsi définis par un rapport de un à un avec les glyphes.

[11] Respectivement très rapprochées des phonèmes français « A », « OU » et « I »

Dans certaines situations, la ligature est obligatoire tel le cas du « Lam-Alef ».

Ces mêmes règles ont été transposées à l'univers électronique au moyen des algorithmes de l'analyse contextuelle qui fonctionnent aujourd'hui de façon réglementaire et adéquate. C'était pourtant le grand problème à résoudre au moment de la transposition des mêmes règles calligraphiques manuscrites à l'univers électronique ; car selon le principe de l'ISO qui stipule un code unique pour un caractère unique, l'ASMO 449 et ses successeurs (ASMO 708=ISO/IEC 8859-6) ne prévoient pas des correspondances de codes avec des glyphes, exception faite des lettres déjà citées : la « Hamza », la « TA » et la « YE » (Tab.31). Ce problème est résolu grâce aux algorithmes d'analyse contextuelle qui agissent à deux niveaux : soit au niveau des périphériques de sortie (Écran, imprimante...) allégeant de la sorte la charge de l'unité centrale de traitement ^[12] (Fig.117)

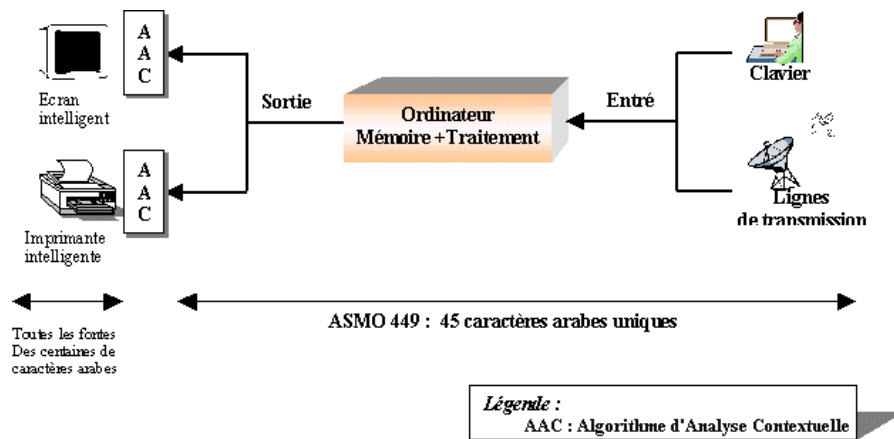


Figure 117 : Le fonctionnement de l'algorithme d'analyse contextuelle au niveau des périphériques de sortie.

Source : Le fonctionnement des outils et des équipements avec le caractère arabe. Mohamed Marayati. In : L'usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.76-117.(EN langue arabe)

Ou bien dans des situations particulières où l'intervention du processeur est indispensable, l'algorithme est confié au processeur central qui se charge de l'exécuter avant de diriger les formes choisies aux unités de traitement des extrants. (Fig.118)

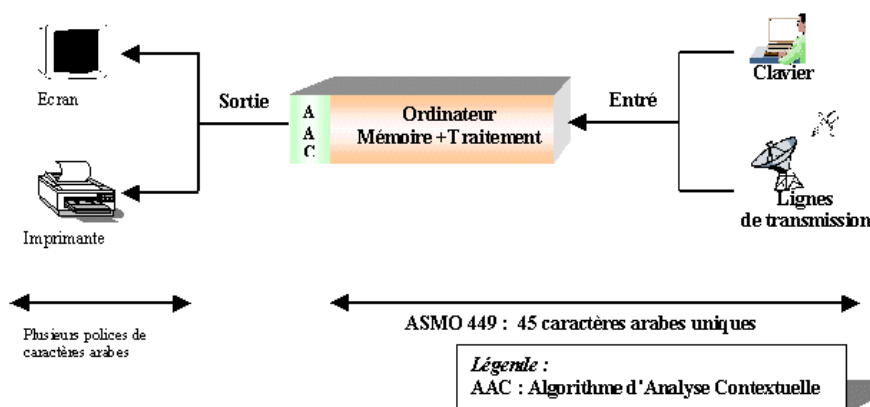


Figure 118 : Le fonctionnement de l'algorithme d'analyse contextuelle au niveau de l'UC.

Source : Le fonctionnement des outils et des équipements avec le caractère arabe. Mohamed Marayati. In : L'usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.76-117.(En langue arabe)

[12] Marayati MOHAMED. Le fonctionnement des outils et des équipements avec le caractère arabe. In : L'usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.76-117.(EN langue arabe)

Parmi les exceptions mentionnées à cette règle, la lettre "Hamza" retient notre intérêt pour la diversité des variantes de glyphes avec lesquels elle est représentée et le silence des textes des normes sur cette décision. L'anomalie que nous observons dans le cas particulier de la « Hamza » est que la sélection de l'un de ses glyphes est très réglementée par la linguistique arabe. Ce qui voudrait dire qu'elle est susceptible d'être traitée selon les algorithmes d'analyse contextuelle de façon stricte et non redondante.

Notre critique serait par conséquent formulée à l'égard de l'adoption de cette variété de glyphes pour un caractère qui normalement jouit d'une réglementation normalisée, sans cas de litige et unanimement reconnue par toute la communauté arabophone depuis les pères fondateurs de la grammaire arabe. Ce qui n'est pas le cas d'ailleurs pour les deux autres lettres « TA » et « YE » qui dépendent beaucoup des pratiques et des usages que l'on en fait dans une zone ou une autre^[13]. Nous proposons d'emblée le constat suivant :

La particularité de la lettre "Hamza" réside dans sa transcription gérée par les règles phonétiques et morphologiques suivantes :

Au début du mot : Si la « Hamza » est articulée par la voyelle "A" (la Fatha) ou la voyelle "OU" (La Dhamma), elle est écrite au-dessus d'une barre verticale (Le Alef). Si elle est articulée par la voyelle "I" (La Kasra), elle est donc représentée sous la barre verticale (Lire la correspondance arabe dans la Fig. de droite à gauche).

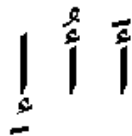


Figure 119 : Les différentes formes des glyphes de la lettre « Hamza » au début du mot

Au milieu du mot : La forme de la « Hamza » dépend de sa propre voyellation ainsi que de celle du caractère qui la précède selon un ordre de poids qui attribue la priorité à la « Kasra », la « Dhamma » puis la « Fatha ». Ces trois voyelles sont respectivement associées aux lettres "YE" dont le glyphe est (ي), "WAW" dont le glyphe est (و) et "ALEF" dont le glyphe est (أ).



Figure 120 : Des glyphes de la lettre « Hamza » au milieu du mot

À la fin du mot : Si la « Hamza » est précédée d'un caractère dont l'articulation est allongée (MADDA), elle s'écrit isolée sur la ligne. Sinon elle est placée au-dessus du caractère relié à l'articulation de la lettre qui la précède (les mêmes règles qu'au milieu du mot dans la figure précédente).



Figure 121 : Des glyphes de la lettre « Hamza » à la fin du mot

[13] H.Mustapha SULEIMAN. Sorting problems of the standard arabic character set. Computer Standards & Interfaces, 1996, n° 18, pp.159-173

Face à ce constat relatif à la première lettre de l'alphabet arabe, la problématique que nous évoquons est la raison de cette solvabilité partielle de la redondance que l'ASMO, dans ses deux normes optimales (449 et 708), a essayé de résoudre intégralement par la définition d'un algorithme d'analyse de contexte comme réponse à la lourdeur des normes transitoires du CODAR (Voir chapitre précédent). Si la transparence (l'opposée de la redondance) dans l'attribution d'un code unique à un caractère unique a été résolue par l'intermédiaire de l'analyse de contexte pour déterminer le glyphe de chaque caractère en fonction de son occurrence au début, au milieu, à la fin d'un mot ou isolé, pourquoi la lettre "Hamza" n'a pas subi le même traitement ? Cette alternative aurait permis de gagner au moins six positions dans la table des codes qui auraient servi pour définir d'autres caractères composés d'un grand taux d'usage dans la langue arabe (i.e. le « Lam-Alef » dont les glyphes apparaissent dans le tableau 25 au niveau des lignes entre 35 et 39).

Nous citons dans ce cas concret relatif au taux d'usage de certains caractères arabes une étude de Merayati^[14] à travers laquelle un prélèvement statistique a été accompli sur un échantillon de 200.000 caractères arabes voyellés et 50 symboles, tous puisés dans des revues, des ouvrages et des journaux scientifiques et saisis en langue arabe standardisée. L'étude a donné lieu à des constatations fort intéressantes qui évaluent les taux d'entropie et de redondance (Tab.36) dans le traitement de la langue arabe.

Taux d'occurrence	Caractère	Taux d'occurrence	Caractère
4.91	ل	.14	ء
2.44	م	6.22	ا
2.08	ن	.98	آ
1.37	هـ	.46	إ
1.14	ة	.06	أ
2.37	و	1.46	ب
.05	ؤ	1.87	ت
.38	ي	.28	ث
.16	ث	.50	ج
2.91	ي	.73	ح
8.77	ق	.36	خ
		1.13	د
12.86	ـ	.40	ذ
7.59	ـ	1.71	ر
3.01	ر	.21	ز
1.67	س	.90	س
.51	ش	.33	ش
.29	ص	.40	ص
.30	ض	.30	ض
.43	ط	.42	ط
.15	ظ	.11	ظ
.01	ع	1.45	ع
.02	ف	.17	غ
5.36	ف	1.11	ف
2.76	ق	.94	ق
15.04	ك	.78	ك

Tableau 36 : Taux de fréquence des caractères en arabe standard

Source : Le fonctionnement des outils et des équipements avec le caractère arabe. Op.Cit. p.87

[14] Merayati MOHAMED. Le fonctionnement des outils et des équipements avec le caractère arabe. Op.Cit. p.86.

$$H_0 = \log(51) = 5.67$$

$$H_1 = \sum_{i=1}^{5012} p_i \log_i = 4.51$$

$$R_1 = 1 - \frac{H_1}{H_0} = 20.46$$

Figure 122 : Formules de calcul de l'entropie dans l'exemple Marayati

Source : *Le fonctionnement des outils et des équipements avec le caractère arabe. Op.Cit. p.87*

Parmi les éléments les plus recensés à travers cette étude, outre les espaces (8.77%) et le manque de voyellation (15.04%), se démarquent la voyelle "Fatha" (12.86%), la voyelle "Kasra" (7.59%) puis la lettre "Alef" suivie de la voyelle "Soukoun" (5.36%), de la lettre "Lam" (4.91%) et de la voyelle "Dhamma" (3.01%)...

En tête de liste apparaissent déjà les deux caractères "Lam" et "Alef" qui, à notre sens méritent une réintégration dans les tables de code comme caractère indépendant avec des variantes de glyphes contextuelles et composées (avec la lettre Hamza).

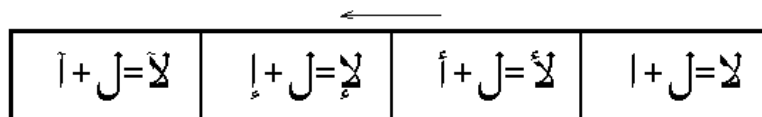


Figure 123 : La structure combinatoire des différents glyphes "Lam-Alef"

D'ailleurs il est actuellement présent sur tous les claviers arabes dans les différentes formes de son occurrence, isolé ou précomposé avec la lettre "Hamza" (Fig.124).

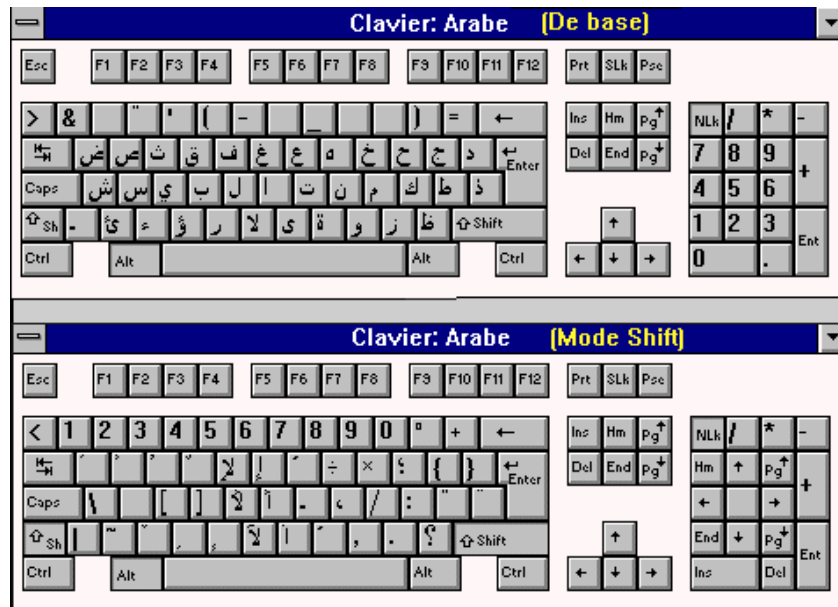


Figure 124 : Le clavier arabe dans ses deux modes de base et Shift

Le caractère "Lam-Alef" a toujours été relégué par les tables de code arabes à l'état de glyphe composé. Aboulhayja^[15] rapporte que ce caractère a été sujet de grandes polémiques au sein des groupes techniques de l'ASMO. Certains spécialistes étaient favorables à son maintien pour des raisons d'ergonomie et de transparence de saisie, d'autres le refusaient pour les considérations suivantes :

- ❑ Conformément au principe de l'unicité de code pour un caractère unique, les caractères "Lam" et "Alef" ne peuvent pas disposer d'une deuxième position dans la table des codes.
- ❑ Les linguistes l'ont simplement rejeté car n'étant pas un caractère de l'alphabet arabe.

À notre sens, nous appuyons la tendance des techniciens qui jugent nécessaire l'intégration du « Lam-Alef » en tant que caractère indépendant dans les tables de code pour les raisons suivantes :

- ❑ De prime abord, l'objectif des concepteurs des tables des codes arabes était de se conformer à l'architecture de la table des codes ASCII pour des raisons de compatibilité, de limitation de nouveaux caractères à coder. En reprenant les mêmes codes de caractères de contrôles, des chiffres et des symboles, il s'agissait d'abord d'un souci d'économie de caractères qui engendre une économie d'espace. Nous ne pouvons donc qu'argumenter que les linguistes arabes, hostiles à l'adoption du caractère « Lam-Alef » dans les tables des codes sous prétextes qu'il n'existe pas en tant que caractère de l'alphabet et que sa présence gênerait tout processus de tri de base, n'est pas solide. Nous montrerons dans les chapitres qui suivent que ces arguments disposent dans la même table des contre arguments. Nous avons déjà vu que des glyphes comme les différentes formes de la « Hamza » ne constituent aucunement des caractères indépendants. Les deux glyphes de la lettre « TA » (ASMO 449 : 4/9 et 4/10 = ASMO 708 : 12/09 et 12/10) sont la représentation non réglementée d'un seul caractère, donc difficile de soumettre à des algorithmes d'analyse contextuelle comme le « Lam-Alef ». Les irrégularités potentielles de tri engendrées par l'introduction du « Lam-Alef » dans les tables de code n'est aussi que supposition non fondée comme nous le verrons dans le chapitre suivant.
- ❑ La deuxième raison pour laquelle nous appuyons l'insertion de « Lam-Alef » comme caractère indépendant dans la table des codes est d'ordre ergonomique. L'objectif d'une technique de codage est entre autres de faciliter la représentation des caractères qu'elle gère par le minimum de redondance et le maximum de transparence. Avec le soutien des algorithmes de l'analyse contextuelle, l'utilisateur du côté de l'interface de dialogue avec la machine, doit voir allégées les complexités du rendu graphique de certains caractères composés et neutres surtout dans un contexte multilingue. Déjà un premier problème de dialogue avec l'interface est posé par la méthode avec laquelle la lettre "Lam-Alef" est codée. Cette lettre apparaît en réalité sous forme d'une série de glyphes composés qui détiennent leurs propres touches sur les claviers arabes (fig.124) sans pour autant disposer de codes appropriés dans les tables des codes. En tapant l'une de ces touches, la forme du "Lam-Alef" en tant que combinaison de deux caractères, qui lui donnent d'ailleurs son nom, s'affiche en bloc comme un seul caractère indépendant. En réalité le rendu en bloc du « Lam-Alef » n'est qu'un effet d'optique car le glyphe comme tel n'existe pas dans la table des codes. Il s'agit de deux caractères « LAM » et « ALEF » qui, en fonction de la rapidité de l'ordinateur dans l'accomplissement de l'algorithme

[15] Ahmed ABOULHAYJA. Les normes et les standards pour l'arabisation de l'informatique. In : L'usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.150-188. (En langue arabe)

d'analyse contextuelle, se joignent pour donner la forme du glyphe « Lam-Alef » dans un laps de temps qui dépasse la perception visuelle de l'être humain. Cette perceptibilité peut toujours être rendu visible si l'on compose le « Lam-Alef » en tapant consécutivement les deux caractères « LAM » puis « ALEF ». Ceci justifie le fait qu'une opération d'effacement arrière du même caractère s'effectuera sur deux temps respectant la composition à deux caractères du glyphe "Lam-Alef" (ou plus s'il est voyellé).

Souvent les utilisateurs ne se rendent pas compte de cette symétrie entre le nombre de caractères nécessaires que le système utilise pour composer le dit caractère (deux caractères) et le nombre de fois qu'il faudrait taper pour l'effacer (Deux fois). La logique dominante sur le mécanisme cognitif de l'utilisateur à ce moment précis est de deux aspects : d'une part la concordance en nombre de frappes entre deux opérations symétriques (saisie à une touche donc effacement à un seul espace arrière) et d'autre part la concordance entre la perception visuelle d'un caractère monobloc et la déduction implicite d'un effacement arrière à un seul espace. Ceci engendre généralement des incohérences de saisie donc des irrégularités de recherche et de tri.

- Une troisième raison à notre avis est que le "Lam-Alef" est généralement confondu avec le "Alef-Lam" en raison de la présence des deux caractères bien qu'en deux modes invertis. Au fait, les "Alef-Lam" constituent l'article défini "AL" en langue arabe. Il ne s'agit pas d'un glyphe comme le "Lam-Alef" mais plutôt de deux caractères indépendants sans ligature. Le problème dans ce cas est que le "LAM" de l'article défini peut créer une ligature avec un "Alef" initial d'un quelconque mot lui succédant (Fig.125). Une ligature sous forme de glyphe "Lam-Alef" aléatoire peut alors être constituée. La problématique de cette situation est que partant d'un principe qui exclue les articles définis de l'ordre de tri des données, la possibilité pour faire la distinction entre le "Alef-Lam" en tant qu'article défini ou en tant que partie intégrante de la chaîne de caractères triée ou recherchée n'est pas largement mise en application par les systèmes actuels de traitement de l'information en langue arabe^[16]. Dans les deux cas de figure, l'aspect cursif de la langue arabe le situe en position rattachée aux caractères qui le suivent. L'absence des diacritiques, phénomène très courant dans l'écriture arabe, complique davantage l'analyse morphologique de la structure des mots et la résolution des homographes (Fig.125).

$$\begin{aligned} \text{(La mère = الأم)} &= \text{الم} + \text{ال} = \text{الام} \\ \text{(Douleurs = الآلام)} &= \text{الام} \end{aligned}$$

Figure 125 : Le Alef-Lam comme syllabe d'un mot ou comme un article défini

Ces remarques constituent, à notre sens, quelques-unes des déboires des algorithmes de l'analyse contextuelle utilisés jusqu'à présent dans le traitement de la langue arabe. Dans l'attente d'Unicode qui apportera certainement la solution radicale à ce problème de traitement de glyphes, il faudrait réactualiser ces algorithmes de façon à répondre à cette situation ainsi qu'à d'autres plus complexes comme la cohérence des opérations du tri alphabétique, la voyellation du texte, la gestion des caractères neutres, la symétrie de la bidirectionnalité des incises, la logique incrémentielle des chiffres et le pivotement des paires de caractères spéciaux. Nous étudierons la majorité de ces aspects dans les paragraphes suivants.

[16] On reconnaît déjà à CDS-ISIS, le logiciel de l'UNESCO la possibilité d'exclure le « Alef-Lam » des opérations de tri et de recherche en saisissant cet article entre les signes inférieur et supérieur (<AL>)

2.2 LE TRI ALPHABETIQUE : LA DIVERSITE DES CONTEXTES ET LA VARIATION DES ALGORITHMES

Nous avons déjà défini dans un chapitre précédent la structure et le fonctionnement algorithmique en guise de préambule à une catégorie particulière d'algorithmes très utilisés dans tous les systèmes d'information automatisés, en l'occurrence les algorithmes de tri. Cette catégorie de tâches est en fait le noyau vital de tous les systèmes d'information manuels ou automatisés qui ciblent l'utilisateur comme aboutissement final de leurs structures opératoires. Or un utilisateur lambda n'a de point d'accès au contenu intellectuel du système d'information qu'à travers un outil ordonné selon un mécanisme de tri alphabétique. Pour ce même utilisateur lambda, le mécanisme de classement intuitif dont il dispose est le résultat d'un référent implicite : les dictionnaires dont disposent toutes les langues. On a tous appris, chacun dans sa propre langue d'origine (et même dans d'autres), l'ordre des lettres de l'alphabet sur lequel les dictionnaires sont ordonnés ou plus particulièrement la façon comment situer les signes utilisés par une langue dans une liste classée. *« Lorsque l'on cherche dans un dictionnaire, on ne dispose en général que de cette connaissance toute simple, et le reste est fait de manière relativement intuitive, sans que l'on ait beaucoup à se préoccuper de la différence entre majuscules et minuscules, de la présence ou non de signes diacritiques, ou même d'autres signes orthographiques (traits d'union inclus dans un mot, apostrophes, espaces, et ainsi de suite) »*^[17].

Quand transposé à l'univers informatique, le tri par ordinateurs retrouve une autre réalité qui n'est plus strictement conforme à la faculté visuelle de l'être humain. Elle est désormais tributaire des contraintes d'ordre technologique plus variées qui ont mis du temps pour s'imposer comme facteurs implicites de tri admis par tout le monde. Nous verrons dans les paragraphes qui suivent comment ce principe de tri par ordinateur n'a pas encore atteint le degré d'uniformité absolue dans tous les systèmes automatisés et pour toutes les catégories d'utilisateurs. On constate encore des anomalies et des incohérences dans le classement alphabétique de corpus textuel unilingues et multilingues. On peut toujours se poser la question sur la raison qui engendre cet état de fait ? La réponse, comme formulée dans la revue « Interface »^[18] est toute simple. *« Avant 1986, jamais personne ne s'était attardé à considérer la différence entre le classement des mots selon la vision humaine et le classement des caractères tel qu'il est possible de le faire très rapidement par une machine. Cela semble incroyable mais c'est la vérité toute simple. Car jusqu'à cette époque, les ordinateurs étaient utilisés beaucoup plus pour traiter des nombres, exécuter des calculs, produire des factures et des chèques. Les listes informatisées de cette époque n'utilisaient que des majuscules, peu de diacritiques sauf dans des cas très particuliers, et le personnel qui manipulait ces listes était entraîné aux problèmes de classement, disposant en général de numéros (comme le NAS) pour retrouver un dossier, puisque retrouver les gens par leur nom s'avérait souvent un problème. Vers les années 1980, toutefois, le traitement de texte a pris son envol de manière accéléré, et l'apparition des ordinateurs personnels a créé le besoin de standards de qualité plus élevés »*.

Il est communément admis que les ordinateurs pratiquent un classement caractère par caractère et mot par mot. Ce qui fait qu'il ne s'agissait palus d'un ordre alphabétique à suivre, mais plutôt d'un ordre où tous les caractères possibles sont considérés en fonction de leurs codes binaires à l'intérieur de la machine. Ceci justifie le fait d'avoir parfois une lettre minuscule « a » triée à la suite d'une lettre majuscule « Z » selon leurs dispositions respectives dans les tables des codes.

[17] La normalisation du classement alphabétique. Interface. Octobre 1997. Vol.8. n°6. URL : <http://www.crim.ca/APIIQ/interfac/>

[18] La normalisation du classement alphabétique. Interface. Op.Cit.

Parfois c'est le cas inverse où l'ordre de priorité est inversé et la lettre majuscule « A » apparaît à la suite suit de la lettre minuscule « z ».

Certains systèmes classaient le terme « vice versa » devant « vice-président »^[19] pour la simple raison que l'espace doit précéder le tiret. Des termes intermédiaires peuvent ainsi s'interposer entre les deux concepts les éloignant l'un de l'autre. Ce qui pourrait engendrer un silence d'information pour un utilisateur qui ne verra pas l'ordre de priorité entre ces signes démunis d'une connotation sémantique et qui pourrait par conséquent se limiter à chercher un terme en fonction de sa position limitrophe à un autre.

L'usage, sans pour autant être guidé par une norme stricte et unique, a canalisé la solution vers un ordre de tri qui prend progressivement racine dans les systèmes de traitement automatisée de l'information. Il s'agit actuellement d'un standard et non d'une norme ; celui d'agencer une liste de caractères graphiques selon l'ordre des symboles (selon l'ordre établi dans les tables des codes), des chiffres (de 0 à 9) et des lettres (de « a » à « z » en minuscule et de « A » à « Z » en majuscule). C'est le principe sur lequel se basent tous les outils (ou presque) documentaires et bibliographiques (catalogues, répertoires, index, dictionnaires, encyclopédies...), tous les outils de Bureautique (les logiciels de traitements de texte, les feuilles de calcul, les systèmes de gestion de bases de données...), tous les produits du génie logiciel. Partout dans ces systèmes, un mécanisme implicite, latent et stratégique agit en continu derrière les différentes couches applicatives des systèmes en question pour donner cette dimension organisée aux données saisies en vrac sans aucun arrangement préalable. Il s'agit des algorithmes de tri.

2.2.1 Les modèles d'algorithmes de tri par échange et par insertion

Les algorithmes de tri obéissent en fait à un grand nombre de stratégies qui diffèrent dans l'ordre des opérations accomplies et la nature même des procédures réalisées. La littérature recense actuellement un grand nombre de ces techniques algorithmiques généralement classées en deux catégories à savoir les modèles de tri par échange (i.e. le modèle de tri bulle, le tri cocktail Chaker, la méthode de Batcher et le tri Quicksort ...) et les modèles de tri par insertion (i.e. par insertion linéaire, par insertion double, tri Shell, tri par arbre binaire...).

Les algorithmes de tri procèdent généralement par un principe d'analogie entre clés de tri qu'ils ordonnent selon des méthodes différentes. Les exemples suivants spécifient les méthodes pratiquées par certains des algorithmes de tri courants.

1. Tri par insertion

1. fonctionnement
 - ❑ insertion à la fin
 - ❑ déplacement vers le début jusqu'à sa position
2. performance
 - ❑ $O(n^2)$
 - ❑ bon pour liste pré-triée, $O(n)$ dans ce cas.

2. Tri Shell

1. fonctionnement
 - ❑ h = distance
 - ❑ examiner les positions $i = h+1, h+2, \dots, n$

[19] Exemple emprunté à la note de référence précédente.

- ❑ trier les positions i et $i-h$, $i-2h$, ...
- ❑ réduire h et recommencer jusqu'à $h=1$
- 2. performance
 - ❑ $O(n^2)$ pour $h=n/2$, $n/4$, ..., 1
 - ❑ $O(n^{3/2})$ pour autre choix de h

3. Tri Quicksort

1. fonctionnement
 - ❑ si l'ensemble est de taille 1 ou 0 arrêter
 - ❑ choisir une valeur dans le "milieu"
 - ❑ partitionner l'ensemble en 2
 - ❑ recommencer sur chaque ensemble
2. performance
 - ❑ $O(n \log n)$ en moyenne
 - ❑ $O(n^2)$ pire cas

Nous détaillons en ce qui suit un modèle de chaque catégorie. Les modèles par « insertion linéaire » et « Quicksort », souvent rencontrés dans des applications courantes de traitement de données textuelles en sont les plus simples. Nous appliquerons ces modèles de tri sur un échantillon textuel choisi en fonction de la variabilité des caractères initiaux par rapport à leurs emplacements dans les différentes zones des tables des codes. Notre objectif est de voir le comportement de ces algorithmes de tri dans la réorganisation d'une liste de chaînes de caractères selon un ordre de tri que nous définissons à titre indicatif à base d'un ordre prioritaire des symboles, des chiffres et des lettres (minuscules puis majuscules).

2.2.1.1 Le modèle de tri par échange « Quicksort »

Ce système est inventé en 1962 par C.A.R.Hoare qui l'a appelé "Quicksort". Ce système permet de positionner un élément à sa place correspondante par analogie avec les 2 éléments qui l'entourent.

Énoncé :

Il faudrait appliquer ce système à un tableau A de N éléments numérotés de 1 à N . Considérons 2 pointeurs i et j qui ont comme valeurs initiales respectives 2 et N . Recherchons, en augmentant i , le premier élément supérieur à A_1 (il marque la limite des chaînes inférieures à A_1) et, en diminuant j , le premier élément inférieur à A_1 (qu'il faudra placer dans le morceau de tableau des chaînes inférieures à A_1). Si $i < j$ il faut échanger A_i et A_j . Il faut ensuite continuer des traitements identiques jusqu'au moment où $i \geq j$. On termine par l'échange éventuel de A_j et A_1 .

Dans notre cas il faut trier les chaînes de l'ensemble A illustré par le tableau suivant :

1	2	3	4	5	6	7	8
1998	réagir	Réagir	[Et.al.]	#2302	?Question	/Auteur	{Champs}

pour arriver par méthode de tri Quicksort à la liste des chaînes triées suivante :

1	2	3	4	5	6	7	8

#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998	réagir	Réagir

Démarche :

Recherchons, à partir de A2 et en nous déplaçant vers la droite, le premier élément supérieur à A1. Il s'agit de A2 équivalent à la chaîne " réagir ".

De même, cherchons à partir de A8 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A8 équivalent à la chaîne " {Champs} ".

1	2	3	4	5	6	7	8
1998	réagir	Réagir	/Auteur	#2302	?Question	[Et.al.]	{Champs}

On les échange

1	2	3	4	5	6	7	8
1998	{Champs}	Réagir	/Auteur	#2302	?Question	[Et.al.]	réagir

On continue à chercher, en nous déplaçant vers la droite, le premier élément supérieur à A1 mais à partir de A3. Il s'agit de A3 équivalent à la chaîne " Réagir ". De même, cherchons à partir de A7 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A7 équivalent à la chaîne " [Et.al.] ".

1	2	3	4	5	6	7	8
1998	{Champs}	Réagir	/Auteur	#2302	?Question	[Et.al.]	réagir

On les échange.

1	2	3	4	5	6	7	8
1998	{Champs}	[Et.al.]	/Auteur	#2302	?Question	Réagir	réagir

On continue à chercher, en nous déplaçant vers la droite, le premier élément supérieur à A1 mais à partir de A4. Il n'y en a pas et i vaut donc 8. De même, cherchons à partir de A6 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A6 équivalent à la chaîne " [Et.al.] " et j vaut {Champs}. Comme i>j, il n'y a pas d'échange mais par contre, on échange A1 et Aj

1	2	3	4	5	6	7	8
?Question	{Champs}	[Et.al.]	/Auteur	#2302	1998	Réagir	réagir

Et l'élément en position 6 est à sa place définitive. Il reste à trier les morceaux de tableau dont les numéros vont de 1 à 5 et de 7 à 8. Commençons par la partie de gauche. Recherchons, à partir de A2 et en nous déplaçant vers la droite, le premier élément supérieur à A1. Il s'agit de A3 équivalent à la chaîne " {Champs} ". De même, cherchons à partir de A5 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A5 équivalent à la chaîne " #2302 ".

1	2	3	4	5	6	7	8
?Question	{Champs}	[Et.al.]	/Auteur	#2302	1998	Réagir	réagir

On les échange.

1	2	3	4	5	6	7	8
?Question	#2302	[Et.al.]	/Auteur	{Champs}	1998	Réagir	réagir

On continue à chercher en se déplaçant vers la droite, le premier élément supérieur à A1 mais à partir de A3. Il s'agit de A3 équivalent à la chaîne « [Et.al.] ». De même, cherchons à partir de A4 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A4 équivalent à la chaîne « /Auteur ».

1	2	3	4	5	6	7	8
?Question	#2302	[Et.al.]	/Auteur	{Champs}	1998	Réagir	réagir

On les échange.

1	2	3	4	5	6	7	8
?Question	#2302	/Auteur	[Et.al.]	{Champs}	1998	Réagir	réagir

On continue à chercher, en se déplaçant vers la droite, le premier élément supérieur à A1 mais à partir de A4. Il s'agit de A4 équivalent à la chaîne « [Et.al.] » et i égale 4. De même, cherchons à partir de A3 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A3 équivalent à la chaîne « /Auteur » et j vaut 3. Comme $i > j$, il n'y a pas d'échange mais par contre, on échange A1 et A_j

1	2	3	4	5	6	7	8
/Auteur	#2302	?Question	[Et.al.]	{Champs}	1998	Réagir	réagir

Et l'élément en position 3 est à sa place définitive. Il reste à trier les morceaux de tableau dont les numéros vont de 1 à 2, de 4 à 5. N'oublions pas la partie de 7 à 8. Commençons par la partie de gauche. Recherchons, à partir de A2 et en nous déplaçant vers la droite, le premier élément supérieur à A1. Il n'y en a pas et $i=2$. De même, cherchons à partir de A2 et en nous déplaçant vers la gauche, le premier élément inférieur à A1. Il s'agit de A2 équivalent à la chaîne « #2302 ». Comme $i > j$, il n'y a pas d'échange entre A_i et A_j mais bien entre A1 et A_j.

1	2	3	4	5	6	7	8
#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998	Réagir	réagir

Occupons-nous de la partie du tableau dont les indices vont de 4 à 5. Recherchons, à partir de A5 et en nous déplaçant vers la droite, le premier élément supérieur à A4. Il s'agit de A5 et $i=5$. De même, cherchons à partir de A5 et en nous déplaçant vers la gauche, le premier élément inférieur à A4. Il n'y en a pas et j vaut 4. Comme $i > j$, il n'y a pas d'échange entre A_i et A_j mais bien entre A4 et A_j. Comme il s'agit du même élément, cette partie est aussi triée.

1	2	3	4	5	6	7	8
#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998	Réagir	réagir

Occupons nous maintenant de la partie du tableau dont les indices vont de 7 à 8. Recherchons, à partir de A8 et en nous déplaçant vers la droite, le premier élément supérieur à A7. Il n'y en a pas

et $i=8$. De même, cherchons à partir de A8 et en nous déplaçant vers la gauche, le premier élément inférieur à A7. Il s'agit de A8 et j vaut 8. Comme $i \geq j$, il n'y a pas d'échange entre A_i et A_j mais bien entre A7 et A_j.

1	2	3	4	5	6	7	8
#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998	réagir	Réagir

Le résultat final est un tri sur le principe annoncé : symboles, chiffres et lettres (minuscules puis majuscules).

2.2.1.2 Le modèle de tri par insertion linéaire

Énoncé :

Nous considérons dans le cas de cet algorithme le vecteur A dont les éléments sont numérotés de 1 à 6. Nous avons réduit à 6 le nombre d'éléments pour éviter la répétition des mêmes tâches accomplies de manière identique sur toutes les chaînes de caractères supplémentaires qui faisait partie de la liste échantillon.

Il est donc prévu, à travers cette méthode de partir de la liste non triée suivante :

1	2	3	4	5	6
{Champs}	[Et.al.]	/Auteur	?Question	1998	#2302

pour aboutir en fin de parcours à l'ordre de tri suivant de la même liste :

1	2	3	4	5	6
#2302	/Auteur	?Question	{Et.al.}	{Champs}	1998

Nous supposons que les éléments numérotés de 1 à k sont déjà triés. Le principe consiste à placer l'élément suivant, à savoir celui de position $k+1$, à sa position parmi les k premiers. Cela se fait en décalant tous les éléments, à partir du dernier élément trié, d'une position vers la droite jusqu'à l'élément inférieur à celui à insérer et de placer ensuite l'élément qui était en position $k+1$ dans le trou ainsi créé.

Démarche :

1	2	3	4	5	6
{Champs}	[Et.al.]	/Auteur	?Question	1998	#2302

Au départ, on considère que la partie triée contient le seul premier élément équivalent à la chaîne « {Champs} » et on désire placer l'élément suivant, à savoir la chaîne « [Et.al.] ».

On déplace la chaîne « {Champs} » vers la droite (puisque classée derrière l'élément « [Et.al.] ») pour placer la chaîne « [Et.al.] » à sa place.

1	2	3	4	5	6
[Et.al.]	{Champs}	/Auteur	?Question	1998	#2302

Nous avons donc les 2 premiers éléments qui sont dans le bon ordre. Il faut placer maintenant le 3^e élément, à savoir la chaîne « /Auteur ». On déplace l'élément « {Champs} » et l'élément « [Et.al.] » (qui sont triés après « /Auteur ») vers la droite pour placer l'élément « /Auteur ».

1	2	3	4	5	6
/Auteur	[Et.al.]	{Champs}	?Question	1998	#2302

Nous avons donc les 3 premiers éléments qui sont dans le bon ordre. Il faut placer maintenant le 4^e élément, à savoir la chaîne « ?Question ». On déplace l'élément « {Champs} » et l'élément « [Et.al.] » (qui sont triés après « ?Question ») vers la droite pour placer l'élément « ?Question ».

1	2	3	4	5	6
/Auteur	?Question	[Et.al.]	{Champs}	1998	#2302

Nous avons donc les 4 premiers éléments qui sont dans le bon ordre. Il faut placer maintenant le 5^e élément, à savoir la chaîne « 1998 ». On ne déplace aucun élément puisque « 1998 » est classé après le 4^e élément, à savoir la chaîne « {Champs} ».

1	2	3	4	5	6
/Auteur	?Question	[Et.al.]	{Champs}	1998	#2302

Nous avons donc les 5 premiers éléments qui sont dans le bon ordre. Il faut placer maintenant le 6^e élément, à savoir la chaîne « #2302 ». On déplace les éléments « 1998 », « {Champs} », « [Et.al.] », « ?Question », « /Auteur » (qui sont considérés postérieurs à l'élément « #2302 ») vers la droite pour placer la chaîne « #2302 ».

1	2	3	4	5	6
#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998

Note: Lors du placement du $k+1^{\text{ème}}$ élément, il faut en moyenne $k/2$ comparaisons pour trouver la place de l'élément dans la séquence déjà triée. Il faudra donc: $(1+2+3+...+N)/2 = [(1+N)/2 * N]/2$ comparaisons. La méthode décrite ci-dessus permet de réduire ce nombre à $N * \log N$ où $\log N$ est le logarithme en base 2.

2.2.2 Les algorithmes de tri : études de cas

Les exigences de collationnement et de tri sont identiques pour tout processus de traitement de données. Une norme d'usage est déjà en vigueur pour organiser tout corpus textuel trié sur la base d'un ordre de priorité des symboles puis des chiffres et enfin des lettres. Un corpus multilingues serait par contre soumis à des irrégularités issues de la particularité de chaque langue traitée.

Le tri alphabétique a été l'un des principes sur lesquels les tables des codes, les pages de codes et les pages de fontes ont été initialement structurées. A part le fait de regrouper les blocs des symboles, des chiffres et des lettres dans un ordre qui correspond aux principes de tri cités auparavant, le fait d'avoir des chiffres ou des lettres, dans leurs zones respectives, disposés de façon contiguë dans les matrices à 7 ou 8 bits explique les intentions initiales des concepteurs

des tables pour faciliter les opérations de tri des données. Or la structuration interne des tables et des pages de codes fait objet de plusieurs interruptions dans l'ordre logique de tri au niveau des deux blocs de symboles (ou caractères spéciaux) et des lettres. Nous porterons notre étude comparative sur les tables des codes qui constituent la source principale des pages de codes et des pages de fontes auxquelles les systèmes informatiques font généralement recours pour gérer le rendu visuel ou imprimé des caractères.

Un détail est cependant à préciser : puisque toutes les tables des codes s'inspirent dans leurs parties gauches de la table ISO/IEC 646, le traitement des spécificités linguistiques françaises (diacritiques) est considéré comme une extension de codes introduite au niveau de la partie droite de la table. La page de codes ASMO 708 inclut ces caractères spécifiques français ainsi que les caractères de traçage des cadres dans les positions vacantes de cette partie droite. Plus particulièrement, tous les caractères alphabétiques accentués sont situés au niveau des deux premières colonnes de la partie droite; soit entre les positions 08/00 et 10/00. Mais leur traitement, comme nous le verrons concrètement est contiguë à leurs équivalents non accentués. Pour des raisons d'allègement des tables des comparaisons dans les chapitres suivants, nous avons assimilés les deux formes à la même désignation de bloc des caractères : le bloc 6 (Tab.37).

Dans la table des codes à 7 bits, les caractères spéciaux occupent quatre blocs de positions regroupées en quatre suites discontinues. Le premier bloc est situé entre les positions 2/1 et 2/15. Le deuxième reprend après les 10 chiffres à partir des emplacements entre 3/10 et 4/0. Le troisième ensemble sépare les zones des lettres majuscules et minuscules au niveau des positions entre 5/11 et 6/1. Le dernier occupe les positions entre 7/11 et 7/14.

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP					
1	SOH	DC1						
2	STX	DC2						
3	ETX	DC3						
4	EOT	DC4						
5	ENQ	NAK						
6	ACK	SYN						
7	BEL	ETB						
8	BS	CAN						
9	HT	EM						
10	LF	SUB						
11	VT	ESC						
12	FF	IS4						
13	CR	IS3						
14	SO	IS2						
15	SI	IS1						

Bloc 1

Bloc 2

Bloc 3

Bloc 4

Bloc 5

Bloc 6

Bloc 7

Légende

Bloc 1 : de 2/1 à 2/15

Bloc 2 : de 3/0 à 3/9

Bloc 3 : de 3/10 à 4/0

Bloc 4 : de 4/1 à 5/10

Bloc 5 : de 5/11 à 6/0

Bloc 6 : de 6/1 à 7/10

Bloc 7 : de 7/11 à 7/14

Tableau 37 : La répartition des blocs de caractères graphique dans la table ISO 646

Étant donnée que la table ASMO 449 est une réplique de la table ISO 646 au niveau de la majorité des caractères avec une transposition des caractères arabes sur les caractères latins qui occupent les mêmes emplacements, et étant donnée que la partie droite de la table ASMO 708

constitue un miroir de la partie gauche au niveau des emplacements des caractères alphabétiques et de certains caractères spéciaux, des ruptures similaires dans les suites des caractères sont aussi engendrées.

En effet, dans une table des codes étendus, des interruptions similaires sont reproduites pour donner les mêmes effets sur des rendus de texte unilingues ou multilingues triés. Des exceptions sont pourtant observées suite aux caractéristiques suivantes particulières à chacune des langues :

- ❑ Inégalité en nombre de lettres entre langue arabe (36 caractères à codes uniques) et langue anglaise (52 caractères à codes uniques) ;
- ❑ Unicité des codes des caractères spéciaux et des chiffres entre les deux langues^[20] (Blocs 1, 2, 3, 5, 7) ;
- ❑ Attribution de codes indépendants réservés aux diacritiques arabes (Bloc 14) ;
- ❑ Non-attribution de codes dans le bloc 15.

	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
00	NUL	DLE	SP													
01	SOH	DC1														
02	STX	DC2														
03	ETX	DC3														
04	EOT	DC4														
05	ENQ	NAK														
06	ACK	SYN														
07	BEL	ETB														
08	BS	CAN														
09	HT	EM														
10	LF	SUB														
11	VT	ESC														
12	FF	IS4														
13	CR	IS3														
14	SO	IS2														
15	SI	IS1						DEL								

	GG	GD
Légende	Bloc 1 : de 02/01 à 02/15 Bloc 2 : de 03/00 à 03/09 Bloc 3 : de 03/10 à 04/00 Bloc 4 : de 04/01 à 05/10 Bloc 5 : de 05/11 à 06/00 Bloc 6 : de 06/01 à 07/10 Bloc 7 : de 07/11 à 07/14	Bloc 8 : de 10/00 à 10/15 Bloc 9 : de 11/00 à 11/09 Bloc 10 : de 11/10 à 12/00 Bloc 11 : de 12/01 à 13/10 Bloc 12 : de 13/11 à 14/00 Bloc 13 : de 14/01 à 14/10 Bloc 14 : de 14/11 à 15/02 Bloc 15 : de 15/03 à 15/10 Bloc 16 : de 15/11 à 15/15

Tableau 38 : La répartition des blocs de caractères graphiques dans une table latin/arabe à 8 bits

Ces interruptions ont nécessité l'intervention de programmes externes pour la mise en forme des produits de sortie selon une disposition qui correspond aux besoins des usagers. Ce sont les algorithmes de tri qui agissent sur le corpus textuel à traiter afin d'agencer son rendu visuel ou imprimé selon des règles prédéfinies. Les systèmes d'information présentent aujourd'hui une grande multitude d'algorithmes qui diffèrent tant au niveau de leurs mécanismes de

[20] Exception faite pour les caractères spéciaux contextuels positionnés aux 10/00, 10/04, 10/12, 10/13, 11/11 et 11/12

fonctionnement interne qu'au niveau de la logique elle-même du principe et de la profondeur de tri qu'ils accomplissent.

Nous démontrerons à travers les exemples suivants la diversité du rendu entre les algorithmes de tri dans les deux environnements microinformatiques les plus courants MS-DOS et MS-Windows dans des contextes linguistiques homogènes (Français à part et arabe à part) et hétérogènes (bilingue arabe/français).

L'expérience que nous menons consiste à définir deux ensembles de mots latins et arabes que nous avons soumis séparément et en combiné à des opérations de tri sous les deux environnements cités. Nous avons échantillonné notre choix des mots de sorte que tous les blocs des caractères illustrés dans les tableaux précédents soient représentés. La comparaison des résultats des opérations de tri nous démontrerait le comportement d'un algorithme et d'un autre dans des contextes différents.

2.2.2.1 Un tri unilingue latin

La confrontation des deux opérations de tri sur les mêmes chaînes de caractères dans un environnement MS-DOS et MS-Windows 95 nous a permis de constater des faits divers dépendants de chacun des contextes suivants :

Algorithme de tri dans un contexte MS-DOS

- ❑ L'algorithme de tri MS-DOS attribue une grande partie de son action de tri à l'ordre préétabli dans les tables des codes. La disposition des chaînes de caractères triées dans la colonne gauche du tableau suivant garde la majorité de l'agencement original des blocs de caractères dans les tables d'origine ;
- ❑ Les blocs 4 et 6 ont été considérés de manière contiguë mais invertie pour respecter un ordre de tri alphabétique des lettres minuscules puis majuscules ;
- ❑ Par la délocalisation logique du bloc 6 au devant du bloc 4, les blocs 5 et 7 ont été rapprochés pour assurer la contiguïté d'une partie des caractères spéciaux.
- ❑ Le bloc 3 réservé aux caractères spéciaux entre 03/10 et 04/00 a gardé sa position d'origine. Il n'a pas été réassigné à la fin de la liste en compagnie des blocs 5 et 7. Ceci a donné comme résultat l'insertion de la chaîne de caractères commençant par le point d'interrogation entre les chiffres et les lettres dans l'exemple de la partie gauche du tableau suivant ;
- ❑ Les lettres sont triées par l'algorithme MS-DOS en tenant compte de la casse et des accents. La priorité est attribuée aux minuscules et aux caractères accentués.

Nous avons également procédé à la validation de ces résultats obtenus par une simple fonction de tri par la commande SORT de MS-DOS en soumettant les mêmes chaînes de caractères à un système de gestion de base de donnée fonctionnant sous MS-DOS, en l'occurrence CDS-ISIS qui a rendu le même ordre de tri.

Environnement MS-DOS	Environnement Windows95

#2302	<i>Bloc 1</i>	#2302	<i>Bloc 1</i>
/Auteur		/Auteur	
1998		?Question	<i>Bloc 3</i>
25 pièces	<i>Bloc 2</i>	[Et.al.]	<i>Bloc 5</i>
?Question	<i>Bloc 3</i>	{Champs}	<i>Bloc 7</i>
réagir		1998	
reagir	<i>Bloc 6</i>	25 pièces	<i>Bloc 2</i>
Réagir		reagir	<i>Bloc 6</i>
Reagir	<i>Bloc 4</i>	Reagir	<i>Bloc 4</i>
[Et.al.]	<i>Bloc 5</i>	réagir	<i>Bloc 6</i>
{Champs}	<i>Bloc 7</i>	Réagir	<i>Bloc 4</i>

Tableau 39 : Comparaison de l'ordre de tri entre la fonction « SORT » de MS-DOS et MS-Word 97

Algorithme de tri MS-WINDOWS 95

Le rendu de MS-Windows 95 réalisé sous MS-Word 97 a donné le résultat affiché dans la partie droite du tableau de comparaison précédent. Cet état de rendu nous permet d'avancer le constat suivant :

- ❑ L'algorithme de tri de MS-Windows applique la nouvelle approche de tri selon l'ordre des symboles, des chiffres et des lettres. Aucune interruption dans cet ordre n'est enregistrée dans l'exemple étudié ;
- ❑ Contrairement à l'algorithme sous MS-DOS, celui de MS-Windows regroupe les trois blocs de caractères 1, 3, 5, et 7 au début de l'ordre logique de tri pour correspondre au principe de l'agencement Symboles, chiffres et Lettres ;
- ❑ Les blocs 4 et 6 attribués respectivement aux caractères Majuscules et Minuscules se trouvent ainsi dans un ordre logique de contiguïté inversée pour respecter la priorité des minuscules sur les majuscules ;
- ❑ L'intermittence entre le bloc 4 et 6 est issue de l'agencement entre les caractères accentués et non accentués. Les diacritiques ont la priorité de tri sur la casse qui intervient en deuxième niveau pour donner la priorité à la minuscule sur la majuscule. Ceci constitue en soi un point de divergence avec MS-DOS qui attribue un poids supérieur aux caractères accentués sur les non accentués. La figure suivante ([Fig.126](#)) illustre l'opposition entre MS-DOS et MS-Windows dans le traitement de la casse et des diacritiques dans les opérations de tri.

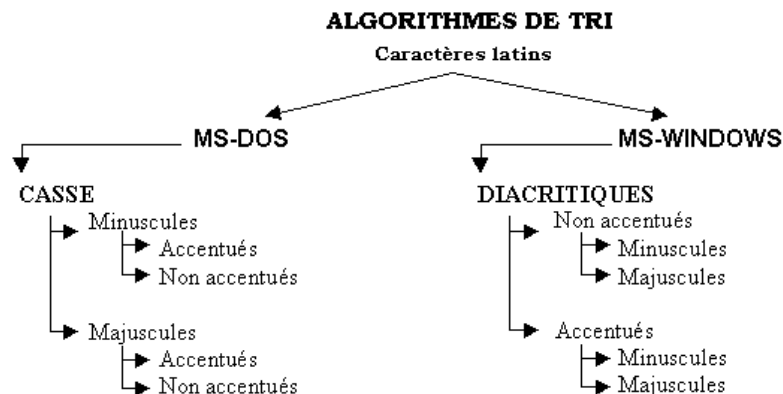


Figure 126 : L'ordre de tri latin dans les deux algorithmes de MS-DOS et de MS-Word 97

Nous avons effectué une série de tests de validation sur ces résultats de rendu qui nous ont effectivement permis de confirmer certains résultats d'une part et de soulever certains cas d'anomalies de l'autre.

Pour la validation de ce rendu par MS-Windows, nous avons soumis la liste des chaînes des caractères à la fonction tri de MS-Access qui a donné les mêmes résultats comme illustrés dans la figure suivante (Fig.127).

1	2	3	4	6	7	8	9	10	11	12
#2302	/Auteur	?Question	[Et.al.]	{Champs}	1998	25 pièces	reagir	Reagir	réagir	Réagir

Figure 127 : Rendu de tri latin sous MS-Access 97

Nous avons par contre soumis la même liste à la fonction de tri de MS-Word en changeant à chaque fois les options de tri pour forcer ou ignorer le respect de la casse. Le résultat a été dans les deux cas identiques.

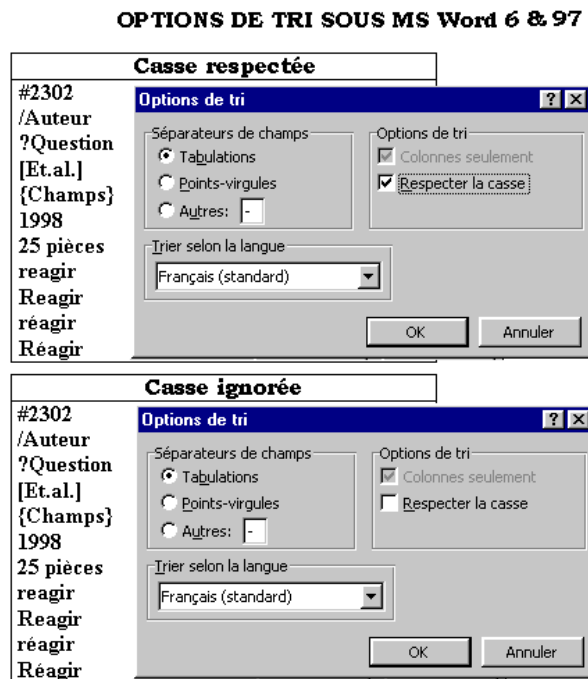


Figure 128 : Anomalie de l'option de respect de la casse dans le tri sous MS-Word 6 & 97

2.2.2.2 Un tri unilingue arabe

Nous avons accomplis les mêmes opérations sur un ensemble de chaînes de caractères arabes choisies en fonction de leurs occurrences dans les différents blocs de caractères comme répartis dans la table des codes étendus. Notre objectif est d'analyser les comportements des algorithmes

développés pour la langue arabe par comparaison à ceux de la langue française dans le tri et l'analyse des chaînes de caractères.

Les expériences ont démontré d'emblée qu'un texte arabe est également trié conformément à la particularité latine dans le contournement des zones d'interruption des blocs des caractères spéciaux et des blocs des caractères alphabétiques. Seulement, l'écriture arabe présente d'autres cas de difficultés issus de la nature même de la langue arabe et des algorithmes de tri mis en œuvre pour les résoudre. Les contextes linguistiques et les modes opératoires suivants présenteront cependant des problématiques de rendu et d'usage que nous essaierons d'analyser au fur et à mesure.

Algorithme de tri dans un contexte Arabic MS-DOS

Le tri de chaînes de caractères arabes sous Arabic MS-DOS applique la même procédure utilisée par les algorithmes de tri latin en se basant en priorité sur le tri préétabli des tables des codes. Comme expliqué auparavant, la structure de la table ASMO 708 est constituée de deux zones de 128 combinaisons binaires chacune. La zone de droite est une réplique « adaptée » de celle de gauche, en l'occurrence la table de code ISO 646. Donc les retombées d'une opération de tri partant de cette structure à 8 bits sont semblables avec toutefois des différences inhérentes aux caractéristiques de l'écriture arabe. La disposition des chaînes de caractères arabes triées dans la figure suivante (Fig.129) garde la majorité de l'agencement original des blocs de caractères dans la table ASMO 708.

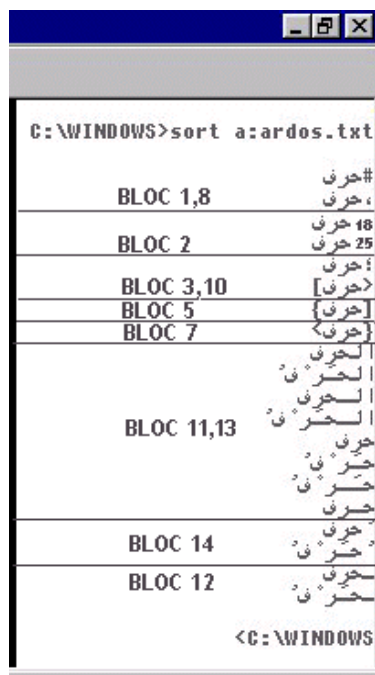


Figure 129 : La fonction SORT pour le tri de texte arabe sous Arabic MS DOS

- ❑ Le bloc 3 réservé aux caractères spéciaux entre 03/10 et 04/00 a gardé sa position initiale avec l'adjonction du bloc 8 qui renferme la virgule contextuelle arabe ;
- ❑ Les blocs 5 et 7 sont traités immédiatement après les blocs 3 et 10 pour garantir un minimum de contiguïté au niveau des caractères spéciaux. Ceci justifie la présence des chaînes de caractères commençant par les accolades et les crochets à la suite de la chaîne commençant par le signe arithmétique « > » ;

- ❑ Le bloc 2 qui regroupe les 10 chiffres n'a pas été délogé pour enlever la rupture au niveau des caractères spéciaux situés dans les blocs 1 et 3. Ce bloc contient les codes appropriés aux deux glyphes de caractères arabes et hindous. Leur affichage dépend d'un choix exprimé par l'utilisateur. Nous reprenons cette particularité dans le chapitre réservé aux chiffres ;
- ❑ Le même bloc 2 trie les chiffres selon un ordre logique et non visuel de saisi. C'est à dire que visuellement, le nombre « 18 » devrait se situer après la chaîne de caractères qui commence par le nombre « 25 » appliquant de la sorte la logique de la comparaison caractère par caractère et mot par mot en commençant par la droite. Or puisque les chiffres sont toujours considérés, d'un point de vue logique ou visuel, dans une directionnalité unique de Gauche à Droite (GAD), l'algorithme de tri commence par confronter les deux chaînes de caractères à partir du premier chiffre situé à gauche pour continuer ensuite avec le premier caractère situé à droite du mot arabe.
- ❑ Le recours au bloc 5 pour faire usage des caractères spéciaux communs aux deux langues, a engendré la contiguïté des blocs 11 et 13 qui constituent les deux parties de l'alphabet arabe ;
- ❑ Le caractère « Tatwil » ou « Kashida » (14/00) du bloc 12 qui sépare les deux parties de l'alphabet arabe est rejetée à la fin de l'ordre de tri arabe après même les diacritiques pour assurer la contiguïté des caractères alphabétiques des blocs 11 et 13 ;
- ❑ Le comportement de l'algorithme de tri des caractères arabes agit différemment de celui des caractères latins sous le même environnement MS-DOS. La suite logique des poids des caractères latins dans une opération de tri est ordonnée selon l'ordre des blocs **1, 2, 3, 6, 4, 5, 7** alors que celle des poids des caractères arabes est ordonnée selon l'ordre des blocs **1-8, 2, 3-10, 5, 7, 11, 14, 12**.

Nous avons également procédé à la validation de ces résultats obtenus par une simple fonction de tri par la commande SORT de MS-DOS en soumettant les mêmes chaînes de caractères arabes au même système de gestion de base de donnée fonctionnant sous MS-DOS, en l'occurrence CDS-ISIS qui a rendu le même ordre de tri.

Algorithme de tri Arabic MS-Word 97

Partant des trois particularités de l'écriture arabe qui la distinguent de l'écriture latine, à savoir :

- ❑ la « Kashida » qui allonge la liaison entre les caractères rattachés pour des raisons de justification à gauche et à droite ou d'espacement des caractères voyellés ;
- ❑ la voyellation des caractères qui peut être nulle, totale ou partielle ;
- ❑ le « Alef-Lam » qui constitue l'article défini soudé à son substantif ;

nous avons opté pour une liste de paires de mots identiques sur lesquelles nous avons apporté des modifications de voyellation, d'insertion de « Kashida » et des préfixes d'articles définis « Alef-Lam » pour voir le comportement des algorithmes de tri dans l'agencement des chaînes de caractères dans des opérations diversifiées de tri. Nous avons procédé au niveau de chaque colonne du tableau suivant à trier les chaînes des caractères selon un croisement pondéré des trois options de tri propres à la langue arabe : « Lam-Alef », Diacritiques et « Kashida ».



Figure 130 : Les options de tri d'un texte arabe sous Arabic Word 6 et 97

L'analyse du tableau des résultats nous permet d'observer les faits suivants :

- ❑ L'algorithme de tri de Arabic MS-Word 97 applique l'approche de tri selon l'ordre des Symboles, des chiffres et des lettres. Aucune interruption dans cet ordre n'est enregistrée dans l'exemple étudié ;
- ❑ Le respect de l'ordre de comparaison caractère par caractère et mot à mot n'est pas pratiqué par l'algorithme de tri de Arabic MS Word 97. Les cellules 13/1 à 16/1 illustrent la non application de cet ordre de comparaison.
- ❑ Contrairement à l'algorithme d'Arabic MS-DOS, celui de Arabic MS-Word 97 regroupe les blocs de caractères spéciaux 1, 3, 5, 7, 8, et 10 au début de l'ordre logique de tri pour correspondre au principe de l'agencement symboles, chiffres et Lettres ;
- ❑ Le seul caractère contextuel (la Kashida) qui reste du bloc 12, lui-même substitué par le bloc 5 pour les caractères spéciaux communs, est repositionné dans une priorité intermittente avec les blocs 11, 13 et 14 conformément aux options de tri activées ;
- ❑ Les blocs 11 et 13 attribués respectivement aux deux parties de l'alphabet arabe^[21] se trouvent ainsi dans un ordre logique de contiguïté pour respecter la continuité de l'alphabet arabe avec ses 36 caractères alphabétiques et ses 8 caractères de voyellation^[22] ;
- ❑ Les lignes 5 et 6 démontrent la source des valeurs binaires des caractères sur laquelle se base l'algorithme de tri sous Arabic MS Windows, en l'occurrence la table ASMO 708 (ISO 8859-6). Dans ces exemples, le point virgule (pivoté verticalement et horizontalement) et le point d'interrogation (pivoté verticalement) devraient, selon la structure des tables des codes, apparaître entre les lignes 2 (/ : barre oblique) et 3 (] : crochet). Ces deux caractères sont en fait des caractères contextuels qui représentent des doublures de leurs équivalents latins. Ils sont placés dans la partie droite de la table des codes au niveau des positions 11/11 et 11/15 respectivement (Voir répercussion dans la partie relative au tri multilingue) ;

[21] La première partie de 12/1 à 13/10 correspond aux emplacements des lettres majuscules dans la partie gauche de la table des codes (=ISO 646) et la deuxième partie de 14/1 à 15/2 correspond aux caractères minuscules de la même partie gauche de la table des codes (=ISO 646).

[22] Les 8 caractères de voyellation 14/11 à 15/2 prononcés respectivement FATHATAN, DHAMMATON, KASRATIN, FATHA, DHAMMA, KASRA, SHADDA et SOUKOUN sont considérés comme des caractères à part entière dont la présence affecte le tri alphabétique des chaînes qu'elles articulent.

K D L	K D L	K D L	K D L	K D L	K D L	K D L	K D L	K - Kashida D - Diacritique L - Alef-Lam
- - -	- + -	- - +	+ - -	- + +	+ - +	+ + -	+ + +	
8	7	6	5	4	3	2	1	حرف
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	1
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	2
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	3
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	4
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	5
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	6
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	7
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	8
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	9
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	10
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	11
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	12
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	13
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	14
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	15
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	16
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	17
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	18
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	19
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	20
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	21
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	22
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	23
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	24
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	25
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	26
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	27
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	28
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	29
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	30
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	31
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	32
حرف	حرف	حرف	حرف	حرف	حرف	حرف	حرف	33

Tableau 40 : L'effet des diacritiques, de la « Kashida » et du « Lam-Alef » dans un tri en langue arabe sous Arabic MS Word 6 et 97

- Une observation générale concerne les diacritiques : prises en compte ou ignorées, les diacritiques ont toujours la priorité absolue entre deux paires identiques dans tous les cas possibles des croisements. À aucun moment, le terme sans diacritiques n'a précédé son équivalent « diacritisé ». Par contre, une « diacritisation » partielle positionne la chaîne partiellement diacritisée entre la diacritisation totale et l'absence totale des diacritiques (analogie entre les cellules 15/1, 16/1, 18/1 et 19/1) ;
- Une deuxième observation générale : quand les trois options sont activées ou ignorées, l'ordre de tri donne la priorité aux diacritiques sur la « Kashida ». Chaque fois qu'il est pris en compte, le « Alef-Lam » est traité comme une suite de deux caractères

alphabétiques et prend donc place dans le corpus de texte en fonction de la position de sa première lettre « Hamza ». Par contre, quand il est ignoré, le tri considère le troisième caractère qui suit les lettres « Hamza » et « Lam ». Les colonnes 1, 3, 5, 7 démontrent que chaque fois que le « Lam-Alef » est pris en compte dans l'opération de tri, ses deux composantes sont triées en fonction de leurs emplacements dans la table des codes. Dans le cas contraire, elles sont ignorées et le tri est basé sur le troisième caractère qui suit (colonnes 2, 4, 6, 8) ;

- Le comportement de l'algorithme de tri des caractères arabes agit différemment de celui des caractères latins dans le même environnement Arabic MS-Windows 95. La suite logique des poids des caractères latins dans une opération de tri est ordonnée selon l'ordre des blocs **1, 3, 5, 7, 2, 6, 4, 6, 4** alors que celle des poids des caractères arabes est ordonnée selon l'ordre des blocs **1, 3, 5, 7, 8, 10, 2**, suivie d'un ordre aléatoire entre **11, 12, 13, 14** en fonction des options de tri sélectionnées. Cette priorité est tributaire des trois options de tri de la langue arabe activées ou ignorées en fonction des besoins particuliers de l'utilisateur.

Cette différence est issue entre autres de l'inversion qu'effectue l'algorithme latin sur les deux blocs Majuscules et Minuscules pour forcer la priorité aux minuscules dans un rendu trié. Puisque la langue arabe ne dispose pas de distinction entre Minuscule et Majuscule, l'inversion n'est donc pas nécessaire.

- Le traitement des options de tri en langue arabe obéit en fin de compte aux règles des priorités illustrées dans la figure suivante (Fig.131) :

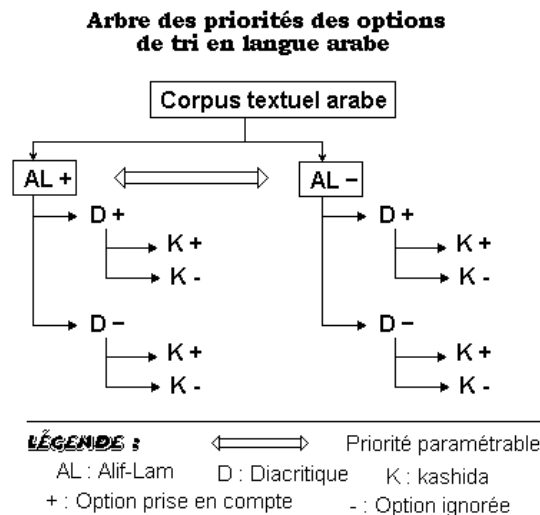


Figure 131 : La priorité entre les options de tri en langue arabe sous MS Word 97

Cette figure illustre les équations suivantes :

- **+ D > + K**

Les diacritiques sont prioritaires sur la « Kashida » quand les deux sont actives. Les cellules 12 et 15 de la colonne 1 le confirment ;

- **+ K > - D - L**

La « Kashida » devient prioritaire quand les diacritiques et le « Lam-Alef » sont ignorées. La colonne 6 le confirme ;

- **+ D > - K**

Les diacritiques sont prioritaires quand la « Kashida » est ignorée ;

- **- D > - K.**

Les diacritiques sont prioritaires sur la « Kashida » même si les deux sont ignorées.

2.2.2.3 Un tri multilingue arabe/latin

Il est mis à l'évidence à travers les exemples précédents que le comportement des algorithmes de tri dans les deux environnements linguistiques arabe ou latin varie sur un grand nombre de points et dépend de plusieurs caractéristiques inhérentes à chacune des langues traitées. Or la question est de savoir comment cette différence est gérée par les mêmes algorithmes quand ils sont supposés intervenir en simultané sur des corpus textuels mixtes.

Pour observer cet état de fait nous avons conduit une série de tests dans les deux environnements standard MS-DOS et Windows à travers une série d'outils courants.

Le tri multilingue sous MS-DOS

La plate-forme des tests est celle d'Arabic MS-DOS (version arabe oblige pour supporter les caractères arabes) configurée à travers la Page de Codes 720 et la Page de Fontes 161.

Test 1 : la fonction « SORT » de MS-DOS

Un premier test par l'éditeur SORT de Arabic MS-DOS a donné les résultats illustrés dans le tableau suivant :

1 !Caractere	21 الحرف
2 حرف	22 الحَرْفُ
3 ,Caractere	23 الحَرْفُ
4 حرف,	24 الحرف
5 25Caractere	25 الحرف
6 25حرف	26 الحَرْفُ
7 :Caractere	27 حرف
8 :حرف	28 حَرْفُ
9 ;Caractere	29 حَرْفُ
10 ;حرف	30 حرف
11 ?Caractere	31 فَرْج
12 ?حرف	32 فَرْج
13 Caractère	33 حرف
14 Caractere	34 فَرْج
15 caractere	35 حرف
16 caractère	36 فَرْج
17 [Caractere}	37 حرف
18 [حرف}	38 حَرْفُ
19 {Caractere]	39 فَرْج
20 {حرف>	40 فَرْج

Tableau 41 : Rendu de tri multilingue sous Arabic MS-DOS en MML

Ce tableau confirme de nouveau le suivi binaire que réalise l'algorithme de tri sous MS-DOS pour ordonner une liste de termes multilingues. Parmi les particularités observées au niveau du rendu du tableau précédent :

- ❑ L'ordre de tri multilingue établit les priorités suivantes :
 - Priorité aux caractères latins sur les caractères arabes dans les deux modes maîtres arabe et latin ;

- Priorité aux caractères arabes sans diacritiques ;
 - Au niveau des caractères alphabétiques arabes, l'ordre de priorité est la lettre, le caractère diacritique et enfin le caractère « Kashida ». Les exemples des lignes entre 33 et 40 établissent clairement cet ordre de priorité.
 - Les paires des lignes 21/22, 25/26, 27/28, 31/32 puis la série de 33 à 40 illustrent cet ordre de priorité.
 - Si cet ordre apparaît inversé au niveau des paires 23/24 puis 29/30 c'est à cause de la priorité du signe diacritique sur la « Kashida ».
 - Priorité de la minuscule sur la majuscule pour les caractères latins ;
 - Priorité des caractères non accentués pour les caractères latins.
- Une fois de plus, l'ordre de tri pour les deux modes et dans les deux écritures est prioritairement basé sur l'ordre binaire des caractères dans les tables des pages de codes et des pages de fontes spécifiées du système. Un algorithme de tri de bas niveau (sous MS-DOS) procède au regroupement des blocs de caractères spéciaux et des caractères graphiques latins et arabes comme spécifié sous la rubrique précédente relative aux opérations de tri unilingues ;
- Des exceptions sont toutefois observées :
- La partie droite de la table des codes, réservée aux caractères arabes est traitée en dernier lieu par l'algorithme de tri multilingue ;
 - L'exception est que les caractères contextuels arabes (10/12, 11/11 et 11/15) qui appartiennent à la partie droite de la table des codes au niveau des blocs 8 et 10, sont traités par l'algorithme de tri la même manière que leurs équivalents latins des blocs 1 et 3. Les exemples des lignes 3 & 4, 9 & 10, 11 & 12 illustrent cette connotation ;
 - Les diacritiques arabes, considérées comme des caractères indépendants, sont traitées conformément à leurs positions dans la table des codes, soit à la suite des lettres non accentuées. Les lignes 27, 28 et 29 confirme la priorité des chaînes sans diacritiques sur les chaînes avec diacritiques, elles-mêmes prioritaires sur la « Kashida ».

Test 2 : Le tri sous le SGBD CDS-ISIS

Le tableau suivant qui reproduit le résultat d'une opération de tri accomplie par le logiciel CDS-ISIS multilingue dans sa version MS-DOS sur un corpus textuel mixte (arabe/latin) démontre l'ordre respecté par le système dans l'agencement du fichier inversé de l'application.

Affichage du lexique	Nom de la Base de Données : TRI
!CARACTERE	الحرف
# الحرف	حرف
,CARACTERE	الحرف
الحرف	حرف
25CARACTERE	الحرف
25حرف	حرف
:CARACTERE	حرف
: حرف	حرف
;CARACTERE	حرف
;حرف	حرف
?CARACTERE	حرف
?حرف	حرف
CARACTERE	فرح
CARACTERE	فرح
[CARACTERE]	فرح
[حرف]	{حرف}
{CARACTERE}	حرف
{حرف}	حرف

Figure 132 : Un rendu de tri sous CDS-ISIS arabisé

Pour analyser ce rendu de tri accompli par CDS-ISIS, les mêmes remarques pour la fonction « SORT » précédente, s'appliquent avec les exceptions suivantes :

- ❑ CDS-ISIS convertit tous les caractères latins en majuscules au moment de la constitution de ses fichiers inversés. Les mots en majuscules et en minuscules sont fusionnés si identiques pour être considérés comme un seul terme ;
- ❑ Les caractères accentués sont gardés intacts, donc affichés comme tels avec une priorité négative par rapport aux caractères non accentués ;
- ❑ CDS-ISIS n'affiche pas les diacritiques arabes mais en tient compte si elles sont saisies à partir du clavier (seule une Kashida de jonction signale leur présence). Ce détail est inhérent aux tables de fontes utilisées. Seules les pages de fontes 162 et 164 permettent d'afficher les diacritiques arabes (excepté le caractère de doublure « Shadda » qui est présent dans toutes les pages de fontes) ;
- ❑ Quand précisé au niveau de la saisie, le « Alef-Lam » peut être ignoré au moment du tri.

Le tri multilingue sous MS-Windows

Le passage à l'univers MS-Windows a constitué, comme démontré à travers les exemples précédents, un gain de qualité dans la définition des algorithmes de tri. Le regroupement par blocs de caractères de même fonction (caractères spéciaux, caractères graphiques...) en est un exemple concret. La page de codes Windows 1256 est un atout supplémentaire pour la restructuration de cet aspect de réaménagement des priorités dans les résultats de tri.

A ce niveau, il nous reste de conclure avec une série de tests que nous avons conduits sous une plate-forme Arabic MS-Windows 95. Pour l'accomplissement de ces tests, nous avons utilisé des outils multilingues dans l'objectif de confirmer cette nouvelle démarche et soulever certains points d'ambiguïté ou d'anomalie qui pourraient résulter de la fusion des corpus textuels arabe et latin.

Test 1 : Arabic MS-Word 97

Nous avons fusionné, comme accompli sous l'environnement MS-DOS, les deux listes arabe et latine sur lesquelles nous avons introduit des modifications pour juxtaposer certaines particularités bidirectionnelles et provoquer certains conflits d'analogie. Par soucis de clarté des résultats de tri, nous avons divisé la liste en deux sous listes. La première est uniquement réservée aux caractères spéciaux et aux chiffres, la deuxième est réservée uniquement aux caractères alphabétiques. Nous avons opté pour cette division car nous avons déjà confirmé le traitement prioritaire de tous les caractères spéciaux avant les caractères alphabétiques. La subdivision n'aura donc aucune répercussion sur l'ordre de tri de l'échantillon entier.

Durant le tri nous avons procédé à une alternance entre les options de tri basées dans ce cas précis sur le respect ou l'ignorance du « Alef-Lam » pour les chaînes de caractères arabes, des diacritiques pour les chaînes de caractères arabes et latins et pour la casse en ce qui concerne les chaînes de caractères latins.

Ces trois critères ont été croisés par trios invertis dans les deux modes maîtres arabe et latin. Nous avons omis la « Kashida » d'une part pour alléger le tableau de comparaison et d'autre part parce que son rôle et ses effets ont été exposés dans le cas du tri unilingue arabe (Voir rubriques précédentes).

Les résultats de la première opération accomplie sur les caractères spécifiques figurent dans le tableau suivant :

Mode Maître Latin		Mode Maître Arabe		
!Caractere	.1	!Caractere	.1	25 حرف
#حرف	.2	#حرف	.2	#حرف
,Caractere	.3	,Caractere	.3	؟حرف
:Caractere	.4	:حرف	.4	؛حرف
؛حرف	.5	:Caractere	.5	؛حرف
;Caractere	.6	;Caractere	.6	{حرف}
?Caractere	.7	?Caractere	.7	{حرف}
[caractere]	.8	{حرف}	.8	:حرف
{حرف}	.9	[caractere]	.9	{caratere>
{caractere>	.10	<حرف}	.10	?Caractere
<حرف}	.11	{caractere>	.11	;Caractere
،حرف	.12	،حرف	.12	:Caractere
؛حرف	.13	؛حرف	.13	[caratere]
؟حرف	.14	؟حرف	.14	,Caractere
25 Caractere	.15	25 Caractere	.15	25 Caractere
25 حرف	.16	25 حرف	.16	!Caractere

Tableau 42 : Tri mixte Arabe-Latin des caractères spéciaux sous Arabic MS-Word 97

Parmi les premières constatations que ce tableau nous permet de mettre en évidence, notons les faits suivants :

- Il est très visible qu'à l'opposé de MS-DOS, l'algorithme de tri sous MS Windows établit une distinction évidente entre un mode maître arabe et un mode maître latin dans la priorité attribuée à une langue ou à une autre. Ceci s'observe à travers les deux colonnes où la priorité linguistique entre les termes arabes et français est inversée en fonction du rapport suivant :
 - Mode maître arabe : le mot arabe suivi du mot latin (i.e. lignes 4 et 5) ;
 - Mode maître latin : le mot français suivi du mot arabe (i.e. lignes 4 et 5).
 - La suite logique du tri des caractères spéciaux sous Arabic MS Word 97 est : **1, 3, 5, 7, 8, 10, 2.**
 - Les lignes 3, 6, 7, 12, 13, 14 démontrent clairement le soubassement ASMO 708 (ISO 8859-6) du système Arabic MS Windows 95. Les chaînes des caractères des lignes 12, 13, 14 sont positionnées à la suite des chaînes qui commencent par des caractères spéciaux conformément à leurs emplacements dans la table des codes ISO 8859-6. Ces trois caractères qui constituent les alternatives contextuelles de la virgule, du point virgule et du point d'interrogation latins (, ; ?) sont traitées par l'algorithme de tri de Arabic MS Word à la suite des blocs 1, 3, 5, 7 au lieu de les traiter, à l'instar de MS DOS, de manière corrélative à leur équivalents latins .
- À notre sens ceci constitue une lacune au niveau de l'algorithme de tri sous Arabic MS Word97 qui ne devrait pas se limiter à un regroupement des blocs des caractères accentués au niveau de la partie gauche uniquement de la table des codes étendus. Puisque l'ordre de tri est fixé selon la priorité Symboles, Chiffres et Lettres, les 4

caractères contextuels spéciaux de la partie droite (le symbole monétaire en plus) de la table des codes étendus devraient être traités comme leurs équivalents latins dans la partie gauche de la table en question afin d'assurer une contiguïté dans le rendu de tri.

- En mode maître arabe, les diacritiques restent prioritaires que ce soit en tenant compte de leur présence ou en les ignorant ;
- Par contre en mode maître latin, les diacritiques sont toujours classées derrière les caractères non accentués ;
- La casse activée donne la priorité aux caractères minuscules dans les deux modes maîtres ;

Mode Maître Latin			Mode Maître Arabe			Légende :
L D C	L D C		L D C	L D C		
- - +	+ + -		- - +	+ + -		
#2302	#2302	.1	#2302	#2302	.1	{Champs}
2302#	2302#	.2	2302#	2302#	.2	[Et.al.]
,regle	,regle	.3	,regle	,regle	.3	#2302
/Auteur	/Auteur	.4	/مؤلف	/مؤلف	.4	Regles
/مؤلف	/مؤلف	.5	/Auteur	/Auteur	.5	/Auteur
;regle	;regle	.6	;regle	;regle	.6	25 pièces
?Question	?Question	.7	?Question	?Question	.7	{تاريخ}
[Et.al.]	[Et.al.]	.8	{تاريخ}	{تاريخ}	.8	{حقن}
{تاريخ}	{تاريخ}	.9	[Et.al.]	[Et.al.]	.9	/مؤلف
{Champs}	{Champs}	.10	{حقن}	{حقن}	.10	1998
{حقن}	{حقن}	.11	{Champs}	{Champs}	.11	règles
،حرف	،حرف	.12	،حرف	،حرف	.12	2302#
؛حرف	؛حرف	.13	؛حرف	؛حرف	.13	?Question
؟سؤال	؟سؤال	.14	؟سؤال	؟سؤال	.14	؟سؤال
1998	1998	.15	1998	1998	.15	1998
1998	1998	.16	1998	1998	.16	Règles
25 pièces	25 pièces	.17	25 قطعة	25 قطعة	.17	25 قطعة
25 قطعة	25 قطعة	.18	25 pièces	25 pièces	.18	regles
regles	Regles	.19	حرف	الحرف	.19	حرف
Regles	regles	.20	حرف	حرف	.20	حرف
règles	règles	.21	حرف	حرف	.21	الحرف
Règles	Règles	.22	الحرف	حرف	.22	عربية
حرف	الحرف	.23	عربية	عربية	.23	عربية
حرف	حرف	.24	عربية	عربية	.24	حرف
حرف	حرف	.25	regles	regles	.25	,regle
الحرف	حرف	.26	Regles	Regles	.26	،حرف
عربية	عربية	.27	règles	règles	.27	;regle
عربية	عربية	.28	Règles	Règles	.28	؛حرف

Tableau 43 : Tri mixte Arabe-Latin des caractères graphiques sous Arabic MS-Word 97

En définitive, l'algorithme de tri multilingue sous Arabic MS Word97 confirme la distinction entre la langue arabe et les langues latines au niveau de l'analogie des chaînes des caractères à trier ainsi que l'ordre de priorité attribué aux caractéristiques d'une langue ou d'une autre. La figure suivante (Fig.133) résume l'ordre de priorité que suit Arabic MS Word97 dans le tri d'un corpus de texte multilingue.

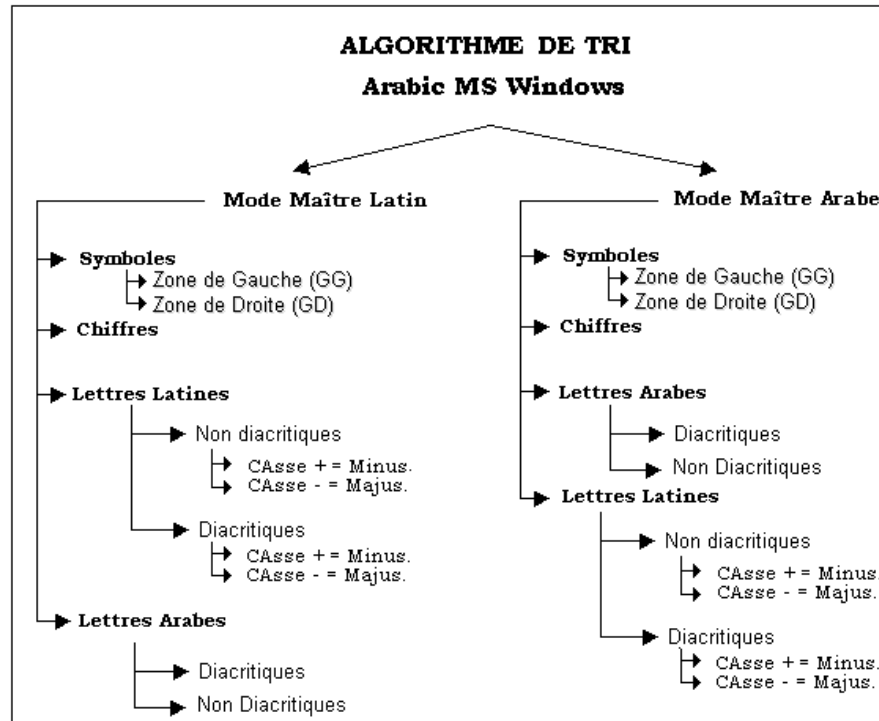


Figure 133 : Critères de tri de Arabic MS Word dans les deux modes maîtres

Test 2 : Arabic MS-Excel 97

Nous avons choisi d'accomplir un test de validation sous MS Excel afin de tester un détail caractéristique qui différencie les logiciels de traitement de texte et les tableurs (ainsi qu'un grand nombre de Systèmes de Gestion de Bases de Données) : la gestion des champs numériques.

Notre soupçon a été justifié par le rendu de tri effectué sous MS Excel97. La figure suivante (Fig.134) met en évidence la différence avec laquelle MS Excel traite les chiffres absolus, soit les chiffres non accompagnés de caractères alphabétiques.

À part le fait qu'ils soient justifiés à droite pour des raisons de calculabilité que nous verrons en détail dans une partie suivante, les chiffres sont triés en priorité absolue créant de la sorte une exception par rapport à l'algorithme de tri sous MS Word.

Une raison possible à ce détail, est sans doute la nature même des applications sous MS Excel dont l'unité d'information majeure est sans doute les chiffres ; raison pour laquelle les concepteurs d'algorithme de tri ont ajouté une petite boucle alternative dans l'ordre des codes sources du programme afin de traiter les chiffres arithmétiques indépendamment de leurs équivalents en alphanumérique. Les cellules A2, A12 et A13 prouvent cette distinction entre les deux sortes de chiffres.

A	
1	Justif.Context
2	1998
3	#2302
4	/Auteur
5	/مؤلف
6	?Question
7	[Et.al.}
8	{Champs}
9	{حقل}
10	بحرف
11	?سؤال
12	25 قطعة
13	25 pièces
	1 ورقة / ورقة 2

Figure 134 : La priorité absolue aux chiffres dans le tri sous MS Excel

2.2.2.4 Le tri multilingue : un premier constat

L'analyse des trois situations de tri pour chacune des langues (en modes isolés ou assemblés) que nous venons de présenter a laissé entrevoir des situations de conformité et de conflits à travers deux niveaux de comparaisons entre la langue arabe et les langues latines, en mode MS-DOS et MS Windows, au niveau des points suivants :

- ❑ Les modes maîtres d'édition ;
- ❑ Présence ou absence des diacritiques ;
- ❑ Sensibilité à la casse ;
- ❑ Gestion des caractères neutres ;
- ❑ Profondeur du tri algorithmique à un ou plusieurs niveaux ;
- ❑ L'effet de la « Kashida » et du « Alef-Lam ».

Nous synthétisons dans la figure suivante (Fig.135) les modes et les méthodes de tri que les algorithmes de tri effectuent pour les deux langues en modes isolés ou combinés dans les deux contextes d'édition que nous étudions, en l'occurrence le contexte MS DOS et le contexte MS Windows.

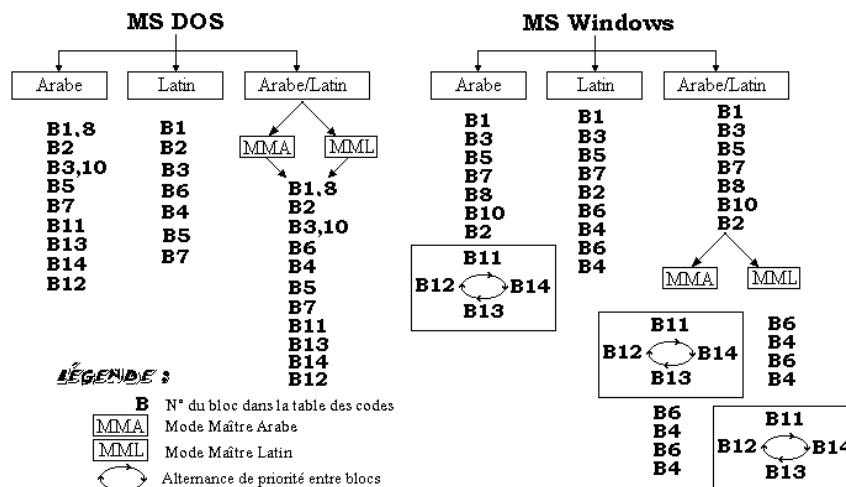


Figure 135 : Modes et méthodes de tri multilingue sous MS DOS et MS Windows

Si nous avons donné toute son importance à la fonction de tri dans les systèmes d'information multilingue, c'est pour exprimer la valeur intrinsèque de cette fonction vitale pour tout système d'information et de communication de l'information scientifique et technique. Tout système vise une optimisation fonctionnelle par la réponse la plus appropriée aux exigences de ses utilisateurs réels et potentiels. Dans la majorité de cas, l'opération de tri s'effectue selon des paramètres qu'un utilisateur particulier spécifie au système conformément à un « ordre alphabétique » conventionnel ou selon des priorités que l'utilisateur choisit en fonction de son référentiel technique, culturel et linguistique.

Les deux cas que nous venons de voir, celui de MS-DOS et de MS Windows, présentent deux contextes où le tri ou l'utilisateur passe d'un mode de travail passif (MS-DOS) à un environnement dynamique (MS-Windows) dans lequel l'utilisateur se trouve impliqué dans la sélection des options de son travail de tri à travers une interface de paramétrage graphique et dynamique ([fig.130](#)).

Dans les deux cas de figures les incohérences de la langue arabe au niveau de l'usage de la « Kashida » et des diacritiques restent permanents car totalement aléatoires et dépendantes des choix souvent inopinés et imprévisibles de l'utilisateur. Il est ainsi recommandé dans des situations professionnelles d'opter pour un prétraitement de tout corpus linguistique arabe avant l'opération de tri. Ce prétraitement consisterait en :

- ❑ Éliminer la « Kashida » du texte pour que l'analogie s'effectue sur les racines réelles des termes. La « Kashida », comme on l'a mentionné auparavant, est utilisée dans deux contextes particuliers. Elle est utilisée pour la fonction de justification gauche et droite d'un paragraphe. Elle est aussi employée selon les goûts aléatoires de l'utilisateur pour aérer une chaîne de caractères ou donner un effet visuel qui relève du goût calligraphique de l'utilisateur ;
- ❑ Harmoniser l'usage des diacritiques pour que ce soit un usage total ou nul. L'usage partiel des diacritiques réduit la conformité de l'ordre de tri partant du fait que les diacritiques sont des caractères indépendants, donc traités comme tels par les algorithmes de tri. Cette précaution revêt une importance particulière car une « diacritisation » partielle est une pratique très courante utilisée pour dissiper des confusions entre des homographes ou des formes grammaticales d'un verbe basées sur l'articulation de la terminaison ;

Dans un cas comme dans un autre, un texte soumis à une opération de tri doit respecter l'une des deux structures morphologiques suivantes basées sur une succession adjacente et réglementée des Consonnes (C), des Voyelles (V) et le cas échéant du caractère « Shadda »^[23] (S) :

CC ...	: Texte sans diacritiques ;
CVCV ...	: Diacritisation totale ;
CSV ...	: Diacritisation totale avec « Shadda » ;

Les diacritiques, la « Kashida » et la « Shadda » (14/00 & de 14/11 à 15/02) sont d'ailleurs fortement déconseillées dans la désignation des noms des fichiers et des dossiers destinés à des

[23] Le caractère « Shadda » (15/01), est un caractère de doublure de consonne. Une consonne double est représentée par le seul glyphe du même caractère superposé du caractère « Shadda ». Son usage est aussi optionnel et sa présence est détectée selon le contexte.

opérations d'échange et de transfert à distance pour les mêmes raisons relatives à l'aspect aléatoire de leur usage.

Or l'ouverture croissante des systèmes d'information et de communication sur des ères linguistiques plus complexes et diverses rend la solution bilingue (arabe et latin) très limitative. Les environnements Client/Serveur qui constituent la majorité, si non la totalité des architectures actuelles des systèmes d'information, pousse le problème du tri à son paroxysme. Les serveurs sont actuellement obligés de répondre à toutes les exigences culturelles de **tous** les clients réels et virtuels, devenant ainsi l'élément responsable du tri multilingue. La solution des tris binaires simples (basés sur les tables des codes) ne sont plus valables culturellement pour aucune langue ni aucune application, sauf les moins exigeantes.

La solution à cette complexité des contextes multilingues et hétérogènes entre déjà dans l'optique de la nouvelle norme de codification sur plusieurs octets, en l'occurrence Unicode et ISO/IEC 10646.

2.2.3 La convergence Unicode/ISO 10646 : une solution interculturelle

Étant donné l'absence d'une plate-forme d'évaluation pratique qui nous fait défaut à ce stade analytique des algorithmes de tri des caractères codés sur plusieurs octets, nous exposerons ci-après un condensé de la littérature publiée par le consortium Unicode concernant ses algorithmes de tri. Nous nous baserons essentiellement sur le rapport technique Unicode n°10^[24] et la section 5.15 du standard Unicode Version 2.0 (anglaise) intitulée « *Sorting and Searching* »^[25].

La particularité multilingue d'Unicode permet de trier toutes les langues individuellement ou en combinaisons de langues variées. La collation de chaînes de caractères Unicode peut prendre en compte les exigences de cultures variées. Une comparaison à plusieurs niveaux permet de trier un texte multilingue en fonction de l'ordre de priorité suivant :

1. les lettres (ou idéogrammes),
2. les caractères accentués,
3. la casse,
4. les caractères spéciaux.

Le mécanisme de tri sous Unicode se base sur une analogie entre une séquence de caractères Unicode et une table de données. Un mécanisme d'adaptation linguistique, permet d'ajuster la table des données pour qu'elle réponde à des particularités linguistiques appelées « Locales ». Ce mécanisme d'ajustement peut intervenir sur les trois premiers niveaux de tri ou sur les quatre à la fois.

À travers une ou deux clés de tri sous forme d'une suite de valeurs 16-bits, une comparaison bit par bit est accomplie pour reorganiser la chaîne des caractères Unicode à trier.

Cette solution fait déjà l'objet d'un projet de norme ISO/IEC 14651 et d'un rapport technique d'Unicode (*Unicode Collation Algorithm*). Cette norme décrit un ordre de classement culturellement correct pour l'ensemble des systèmes d'écriture couverts par le jeu universel de caractères ISO/IEC 10646. Cette norme couvrira la plupart des écritures alphabétiques (les

[24] Lire détails sur URL : <http://www.unicode.org/unicode/reports/tr10/index.html>

[25] The Unicode Standard, Version 2.0. Addison Wesley Longman, 1996.

écritures latines, cyrillique, grecque, arabe, géorgienne, arménienne), ainsi que les écritures idéographiques (i.e. le chinois) et syllabiques (i.e. le hangul coréen, les kanas japonais). Il est prévu de couvrir tous les systèmes d'écriture contemporains et historiques (i.e. hiéroglyphique).

L'un des grands apports attribués à cette norme est d'aller au-delà du simple classement pour pouvoir normaliser également la méthode de comparaison entre deux rubriques. Cette possibilité permettrait non seulement d'élaborer un classement alphabétique, mais permettrait aussi de déterminer les équivalences linguistiques et alphabétiques de recherche. Si l'on interrogeait un moteur de recherche sur Internet par le concept « laser », le système pourrait restituer, selon la convenance de l'utilisateur, toutes occurrences des concepts « LASER », « LAZER », « laser », « lazer ». Par contre, il serait également possible de retrouver uniquement le mot « là » avec l'accent grave sur la lettre « a » si l'on voulait faire une recherche pointue afin d'éviter la récupération d'un taux élevé de bruit par un terme vide comme l'article défini « la ». Ce sont là (!) quelques problèmes qui seront résolus par la norme ISO/IEC 14651, dans une interface de programmation que l'on pourra généraliser à tous les systèmes informatiques de l'avenir.

Des produits multilingues commencent déjà à appliquer cette nouvelle norme universaliste qui ne s'applique à aucune culture en particulier, mais qui sert de point de départ pour *l'adaptation* du classement à une ou à plusieurs profils culturels reconnus dans le contexte de l'internationalisation du traitement multilingue sous le concept de **locales**.

Une locale est la portion dans un environnement utilisateur tributaire des facteurs langue, pays et conventions culturelles d'usage. Elles ne sont en fait qu'un ensemble de paramètres de préférences linguistiques de l'utilisateur présentés sous forme d'identifiant de codes de langues (LCID = Language Control IDentifier) paramétrables et interchangeable durant tout le processus de travail.

Un LCID est une valeur utilisée pour représenter et définir de manière unique et univoque une locale en utilisant les trois paramètres d'un identifiant de langue principale ; d'un identifiant de langue secondaire et d'un identifiant de mode de tri pour les cas où une langue aurait deux méthodes de tri (le japonais en a le mode JIS et le mode Unicode). Les locales peuvent également inclure un grand nombre de spécificités systèmes comme les noms des mois, des jours, des pays, des langues, des valeurs de pages de codes, des formats standards pour les dates, le temps et la monnaie...

Dans un système multilingue, les locales servent de plates-formes nationales pour permettre à un système basé sur la norme ISO 14651 de tri multilingue de correspondre à des exigences culturelles locales ou nationales. Les locales qui constituent ainsi des conventions culturelles pour le traitement des langues nationales dans un contexte multilingue font l'objet de la norme ISO 14652.

La nouveauté d'un tri sous Unicode est qu'il « *innove et prolonge le modèle de locale traditionnel dans lequel tout traitement sensible à la culture (tel le tri) s'effectue dans le cadre d'une locale installée qui précise les conventions culturelles en matière d'informatique. Le tri permet de spécifier plusieurs locales, afin que dans un fichier qui contient plusieurs langues, celles-ci puissent être triées selon leurs propres conventions* »^[26].

[26] Voir l'URL : <http://www.Alis.com>

Le concept des locales est encore à ses débuts bien que Microsoft en a déjà fait objet dans ses programmes de support des langues nationales (NLS = National Language Support) implantés dans ses différentes versions de son système Windows 32 bits.

Nous avons analysé certains produits de Bureautique basés sur un codage Unicode. La fonction de tri, si elle existe, n'y est pas encore très développée. Nous avons essayé dans ce sens le logiciel de traitement de texte « *Accent Working Model* » de la société Accent Software International Ltd (GB, EUA) qui ne dispose pas encore d'une fonction de tri.

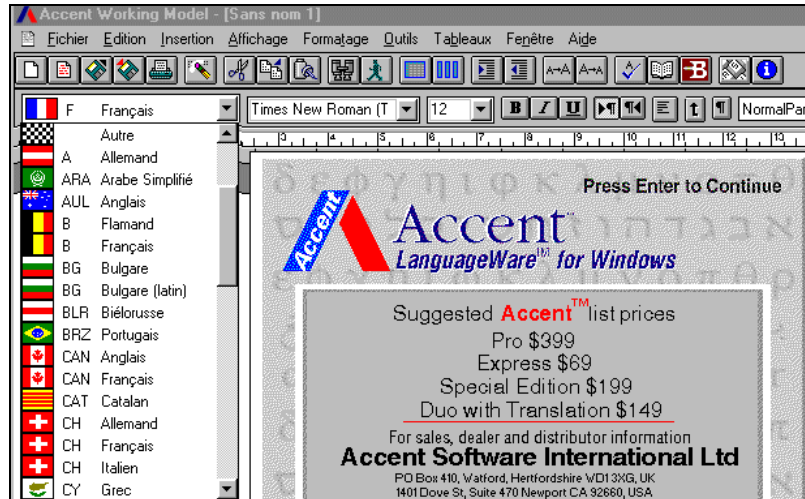


Figure 136 : La multilinguisme des nouvelles applications de traitement de texte chez Accent Software Ltd.

Nous avons également observé les logiciels de traitement de texte « Universal Word »^[27] de WYSIWYG Corporation (EUA) qui n'intègre la fonction de tri qu'au niveau de la gestion des opération de fusion et publipostage^[28].

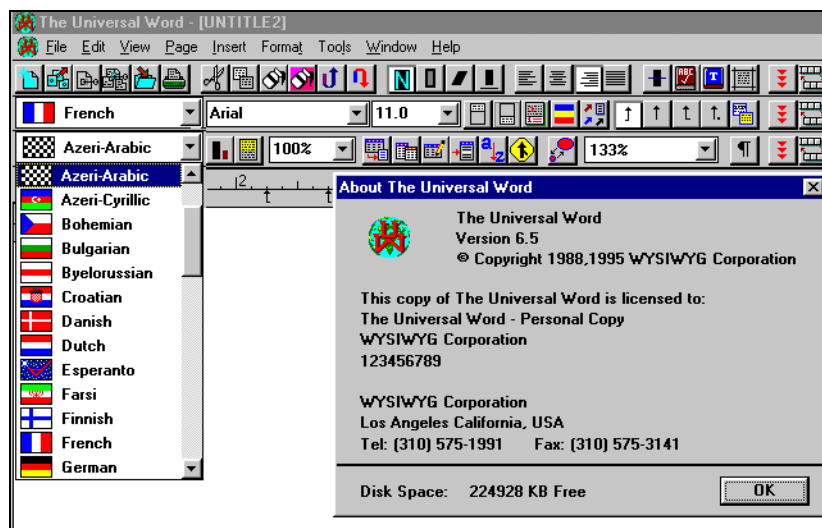


Figure 137 : Le traitement multilingue sous Universal Word

[27] Voir l'URL : <http://www.wysiwyg.com/>

[28] Nous n'avons pas pu procéder à un test de cette fonction car elle est désactivée dans la version de démonstration que nous possédons du dit produit.

Nous reviendrons à ces produits dans le chapitre suivant relatif à l'analyse des interfaces multilingue Homme-Machine pour analyser d'autres aspects du traitement multilingue des données.

2.3 LA BIDIRECTIONNALITE

La bidirectionnalité de l'acte scriptural est l'une des sources de difficultés majeures entre les écritures arabe et latines. Le texte arabe étant écrit de droite à gauche (DAG), une mixture de texte arabe et latin devrait donc gérer les deux directionnalités à la fois sous forme de blocs de texte isolés ou fusionnés. Le deuxième cas présente plus de complexités car la bidirectionnalité ne se limite pas à gérer deux sens d'écriture dans deux ou plus parties différentes d'un même document, mais plutôt à gérer la même bidirectionnalité sur la même ligne. Trois aspects de complexités d'un texte mixte et entrelacé sont alors observés : les incises et les bris, les caractères neutres et les chiffres.

Avant d'aborder ces trois aspects du point de vue de leur répercussion sur l'ergonomie de l'interface bilingue Homme-Machine, un aperçu sur le traitement interne des caractères multilingues permettrait de comprendre le mécanisme avec lequel les systèmes traitent les données textuelles multilingues que ce soit au niveau de leur stockage interne ou du point de vue de leur rendu visuel et imprimé.

2.3.1 Le traitement logique (Backing store) et la conversion visuelle d'un texte Bidi

Le texte saisi est en réalité stocké dans les registres de la mémoire tampon (Backing store) dans un mode séquentiel logique. L'ordre logique de stockage correspond à l'ordre dans lequel les caractères sont saisis. Cette méthode est adoptée d'une part pour des raisons de linéarité dans la gestion de l'espace mémoire et d'autre part pour faciliter l'une des tâches d'édition les plus courantes et non moins complexes comme les combinaisons « Copier/Couper et Coller ». En effet, pour faciliter ces opérations et les rendre transparentes pour le système interne, les caractères gagneraient à être alignés de façon contiguë en un seul bloc. La conversion de l'état de stockage logique et séquentiel des caractères à l'état de rendu visuel ou imprimé des glyphes constitue l'un des grands problèmes du traitement de la langue arabe. Pour optimiser la rentabilité et la transparence des applications multilingues, cette conversion se fait à travers des routines de restitution de texte de bas niveau. Il s'agit des « algorithme bidi ».

Dans l'exemple suivant, les caractères dans les cases de 9 à 15 montrent le terme arabe stocké en mémoire dans une disposition logique séquentielle de gauche à droite (GAD) conformément à la direction du mode maître latin et l'agencement des caractères dans la table des codes en usage, en l'occurrence la table ASMO 708 (ISO/IEC 8859-6).



Figure 138 : La conversion entre les deux modes logique et visuel en mode maître latin

L'inverse est aussi vrai et le « Backing store » s'effectue de droite à gauche en mode maître arabe. Le rendu visuel ou imprimé est par contre adapté par l'algorithme Bidi qui inverse la disposition des caractères pour leur donner une orientation conforme au mode maître actif. (DAG ou GAD) et une apparence de glyphs conformément à une analyse contextuelle stricte.



Figure 139 : La conversion entre modes logique et visuel en mode maître arabe

Si le rendu dans les deux cas est identique c'est suite à l'inversion de saisi des deux concepts opposés. Si la même suite de caractères est saisie dans le même ordre dans les deux modes maîtres d'édition elle donnera deux rendus opposés car l'algorithme Bidi aurait agit doublement par l'inversion des chaînes de l'incise et l'inversion de sa position par rapport au texte hôte.

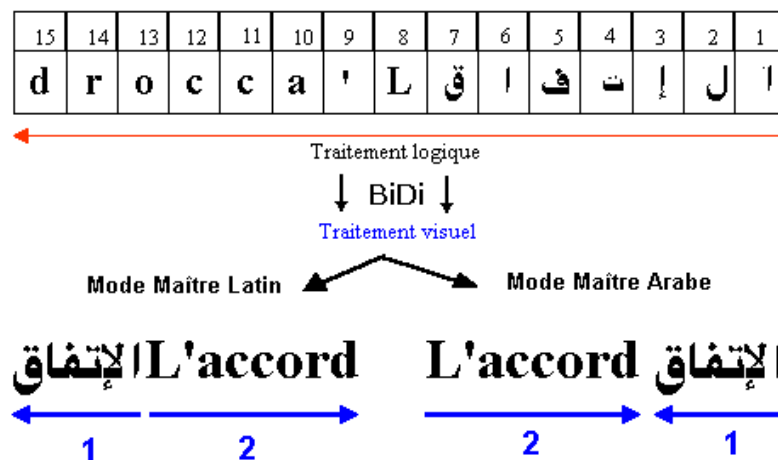


Figure 140 : Inversion de position des chaînes opposées selon le mode maître actif

La conversion entre traitement logique et rendu physique est également perceptible à travers la fonction de sélection de texte et le mouvement du curseur. Encore une fois, un mode maître est l'opposé de l'autre. Nous nous contenterons d'œuvrer sur le modèle du mode maître latin essentiellement avec invocation du mode maître arabe quand nécessaire.

Une sélection des chaînes de caractères de l'exemple précédent s'accomplit dans les deux modes à la fois : logique en interne et visuel en externe.

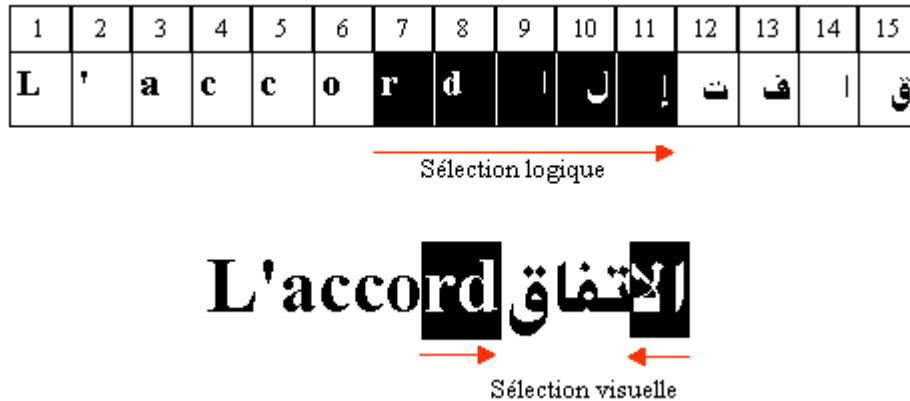


Figure 141 : Modes de sélection interne (logique) et externe (visuel) en mode maître latin

Plus compliqué encore est le cas de la sélection effectuée sur une chaîne de caractères mixtes incluant des chiffres.

L'une des constantes du traitement informatique des données est l'unicité directionnelle des chiffres de gauche à droite (GAD) dans les deux modes maîtres arabe et latin. Nous aborderons ce détail avec beaucoup plus d'approfondissement sur les points techniques et surtout culturels dans les paragraphes suivants. D'emblée, cette unicité directionnelle des chiffres est en faveur du texte latin qui constituerait ainsi un prolongement directionnel uniforme de l'acte graphique ou de traitement. Ce qui n'est pas le cas pour le texte arabe qui doit dans ce cas de figure traiter les chiffres d'une manière opposée à sa direction normale de droite à gauche (DAG). Ainsi, dans l'exemple suivant édité en mode maître latin, les chiffres gardent leur unicité directionnelle. Une opération de sélection serait plus compliquée du côté du texte arabe que celui du latin.



Figure 142 : Unicité directionnelle GAD (LTR) des chiffres en MML

L'inversion vers le mode maître arabe ne fait que compliquer les choses car à part le maintien de la même complexité de la suite lettres et chiffres arabes, l'ordre sémantique de la phrase lui-même pivote.

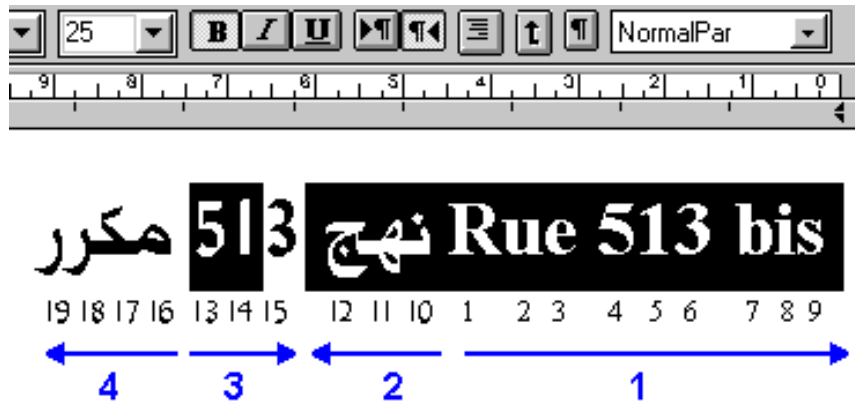


Figure 143 : Unicité directionnelle GAD (LTR) des chiffres en MMA et inversion sémantique

Cet exemple est non conforme à l'édition Bidi en mode maître arabe car le texte Bidi de chaque mode maître doit normalement commencer dans la langue qui lui correspond et dans une direction qui lui convient. La forme physique (visuelle) d'un texte Bidi est donc strictement dépendante de sa représentation interne ainsi que du mode maître d'édition dans lequel il est saisi. Le mouvement non linéaire du curseur et de la sélection, la réorganisation sémantique des bris dans les incises et l'unidirectionnalité des chiffres sont des détails très latents et souvent imperceptibles par l'utilisateur profane qui se trouve ainsi facilement perturbé par l'ordre des actions entreprises par les algorithmes qui gèrent les différentes phases d'un texte Bidi.

Le mouvement du curseur qui constituent le repère visuel de l'utilisateur sur le rendu visuel peut cependant être paramétré selon l'un des deux modes en question. Certaines interfaces graphiques courantes sous MS Windows proposent cependant les deux modes à l'utilisateur.

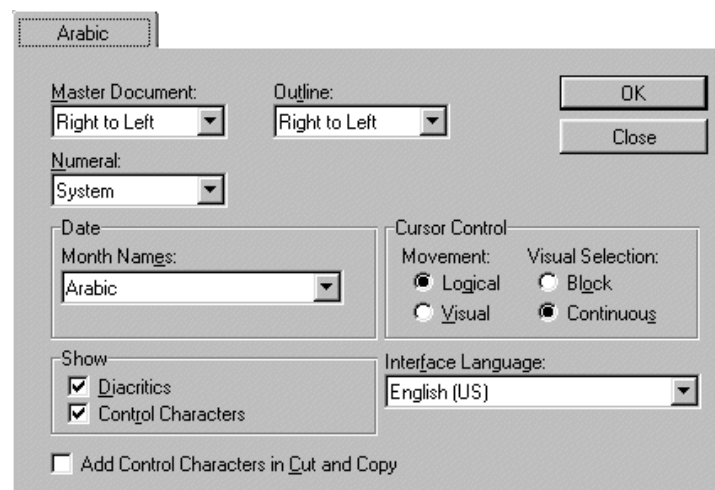


Figure 144 : Le paramétrage du curseur en modes logique ou visuel sous MS Word97

2.3.2 Les incises et les bris de lignes

Pour comprendre la raison de l'adoption de cette méthode d'inversion entre mode logique et mode visuel, il suffit d'imaginer la difficulté d'une situation de bris d'une incise dans une langue opposée stockée en mode visuel. Pour assurer une lecture naturelle, les premiers mots de l'incise

doivent rester sur la première ligne avec la première partie de la phrase principale, mais ils viennent en dernier en ordre visuel.

En ordre logique, il suffit de briser la chaîne de caractères au point approprié et de laisser travailler l'algorithme bidi sur chaque ligne résultante. Les deux exemples suivants illustrent l'importance du traitement logique pour une restitution visuelle adéquate.

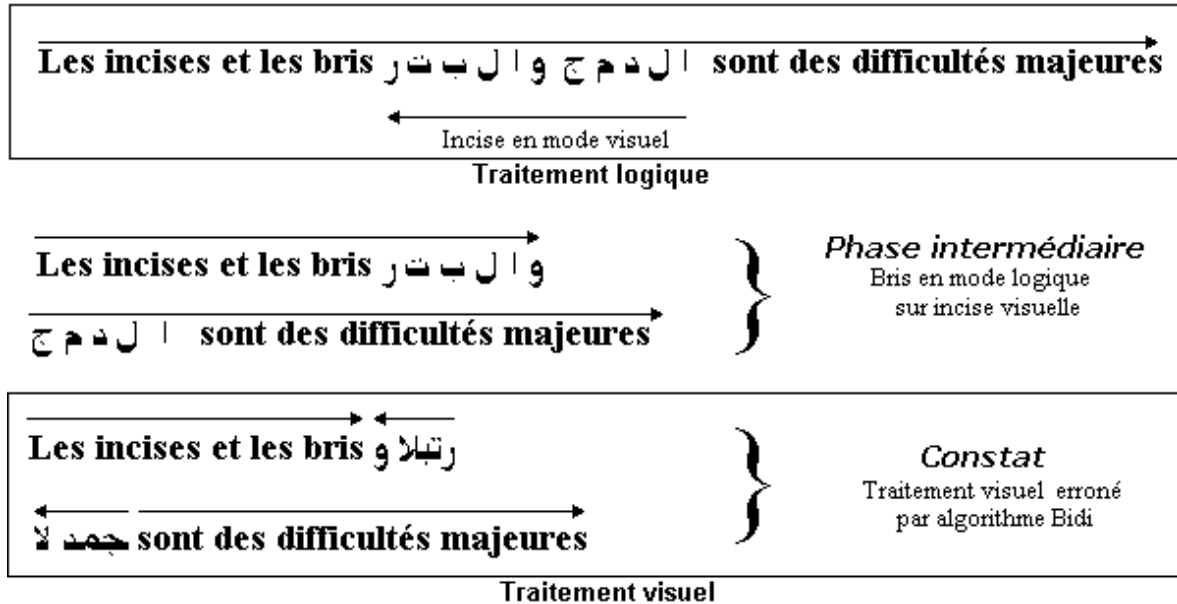


Figure 145 : Incise et bris en modes logique et visuel erronés

Il est vrai aussi que la technique de l'inversion du mode de traitement logique évite au système de faire un saut de mémoire vers la partie droite pour commencer une chaîne de caractères arabes. En plus du fait qu'il n'est nullement indiqué au préalable le nombre d'espaces mémoire qu'une incise va occuper avant d'arriver à la limite du texte latin opposé, la linéarité dans la gestion des adresses mémoire évite la création des creux et des zones tampons non allouées.

Une interrogation reste cependant plausible dans une pareille situation. On pourrait toujours faire en sorte qu'un algorithme adapté puisse gérer *en mode logique* une incise d'une directionnalité opposé en utilisant un mécanisme d'injection. Un bloc de caractères d'une langue principale se termine en mode décalage en un point précis de la zone mémoire à partir duquel la deuxième langue est insérée en mode injection ou glissement. Le curseur se fige et les caractères imprimés sont injectés et glissés vers la direction de l'écriture qu'ils représentent. Le résultat aurait été un « Backing store » logique identique au rendu visuel ou imprimé. On aurait pu ainsi éviter les complexités inutiles d'un algorithme Bidi.

À notre sens, la réponse à cette hypothèse passe obligatoirement par l'analyse de la structure purement électronique d'un système informatique et du fonctionnement des registres d'une zone mémoire ainsi que des techniques de leurs allocations ; perspective qui dépasse la portée de notre travail actuel.

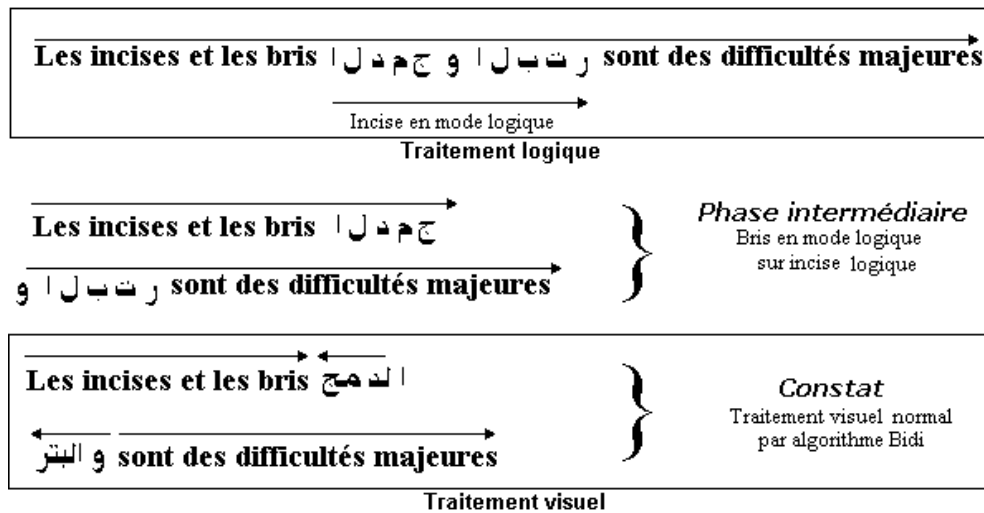


Figure 146 : Incise et bris en modes logique et visuel correctes

Les deux cas d'incises et de bris présentés sont basés sur un contexte d'édition en mode maître latin qui coïncide d'ailleurs avec l'ordre dans lequel les caractères sont classés dans la table des codes étendus. Or l'inverse est aussi vrai car un stockage logique des données dans les registres de mémoire s'accomplit dans l'ordre séquentiel de leur saisie. Une édition en mode maître arabe suivrait un stockage de droite à gauche.

Par mode maître d'édition l'on sous-entend que le mode d'édition par défaut est dans une langue et que l'autre langue est introduite en mode insertion (donc en une directionnalité opposée). Ceci donne en fait deux modes d'édition possibles :

- Un mode maître latin de gauche à droite avec insertion en arabe de droite à gauche : Ce mode correspond à une bidirectionnalité opposée qui évolue de gauche à droite tout en juxtaposant les terminaisons des blocs multilingues.

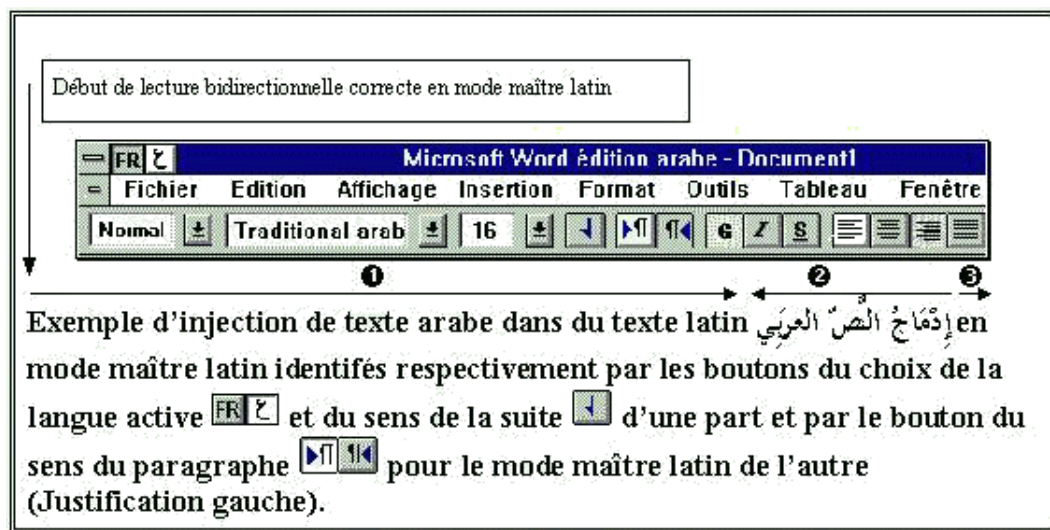


Figure 147 : Incise arabe correcte en mode maître latin

La lecture d'un texte mixte dans ce mode s'effectue à partir du point le plus à gauche de l'écran et évolue vers la droite. Au niveau d'une incise arabe, la lecture balance vers la droite

du texte arabe et évolue vers la gauche jusqu'à la fin de l'incise. La figure suivante (Fig.148) illustre le mouvement des yeux dans la lecture d'un texte mixte dans ce mode précis.

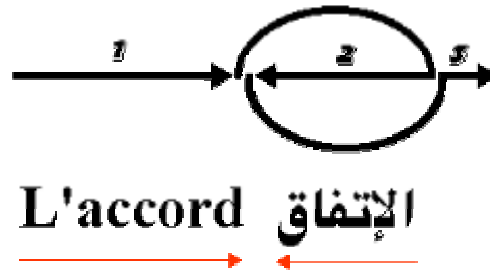


Figure 148 : Le sens de lecture d'une incise arabe en mode maître latin

- Un mode maître arabe de droite à gauche avec insertion en latin de gauche à droite : Il s'agit du même principe mais dans un sens de droite à gauche. Dans l'exemple suivant, le bris de la première ligne ayant eu lieu au niveau du texte arabe qui constitue le mode maître actif, a effectué un retour à la ligne à droite conformément à l'orientation du texte et du mode maître d'édition à la fois.

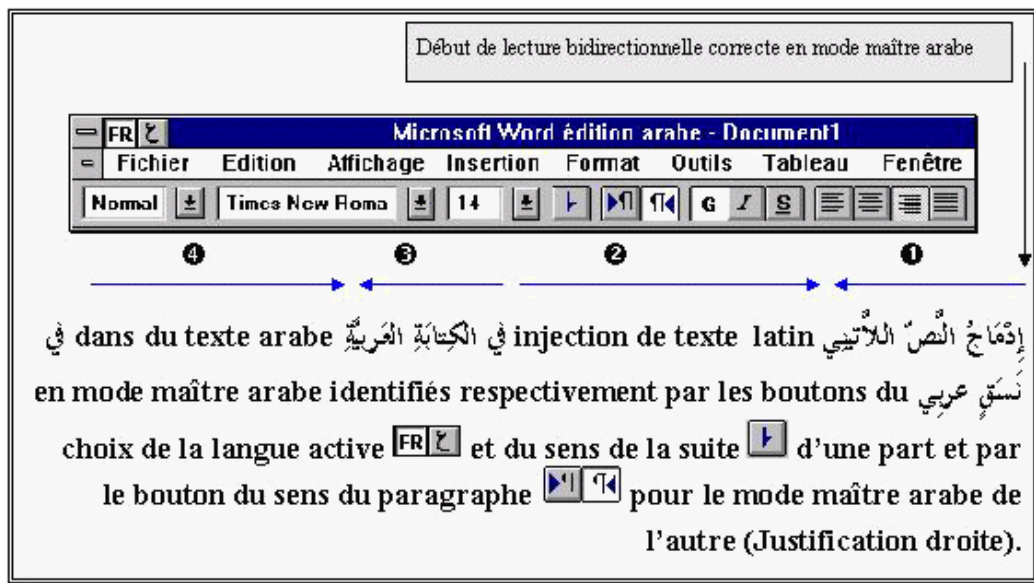


Figure 149 : Incise latine correcte en mode maître arabe

La lecture d'un texte mixte dans ce mode s'effectue à partir du point le plus à droite de l'écran et évolue vers la gauche. Au niveau d'une incise latine, la lecture balance vers la gauche du texte latin et évolue vers la droite jusqu'à la fin de l'incise. La figure suivante (Fig.150) illustre le mouvement des yeux dans la lecture d'un texte mixte dans ce mode précis.

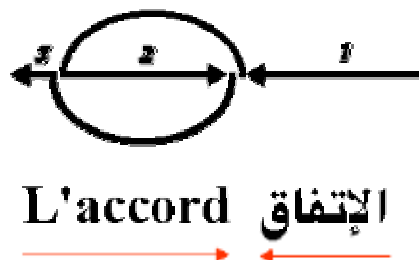


Figure 150 : Le sens de lecture d'une incise latine en mode maître arabe

Ces deux exemples constituent les deux modes d'édérations multilingues normaux ou la directionnalité respecte à la fois l'orientation standard de la langue et du mode maître d'édition. Cependant, si nous appliquons une combinaison pondérée à ces 4 cas de figures (Mode maître arabe, mode maître latin, orientation gauche-droite et orientation droite-gauche), deux autres combinaisons demeurent toujours possibles bien que complètement inadéquates :

- Un mode maître latin de droite à gauche avec insertion arabe de gauche à droite : Ce cas de figure présente une opposition de blocs linguistiques dos à dos. Ce qui constitue une opposition d'écriture incorrecte. Dans l'exemple ci-dessous l'apparition du bloc arabe 1 en début de ligne éditée en mode maître latin est le signe de l'incohérence de la direction du paragraphe avec le mode d'édition global. Le schéma d'agencement des blocs linguistiques n'est donc pas conforme à la logique évolutive du sens de l'écriture.

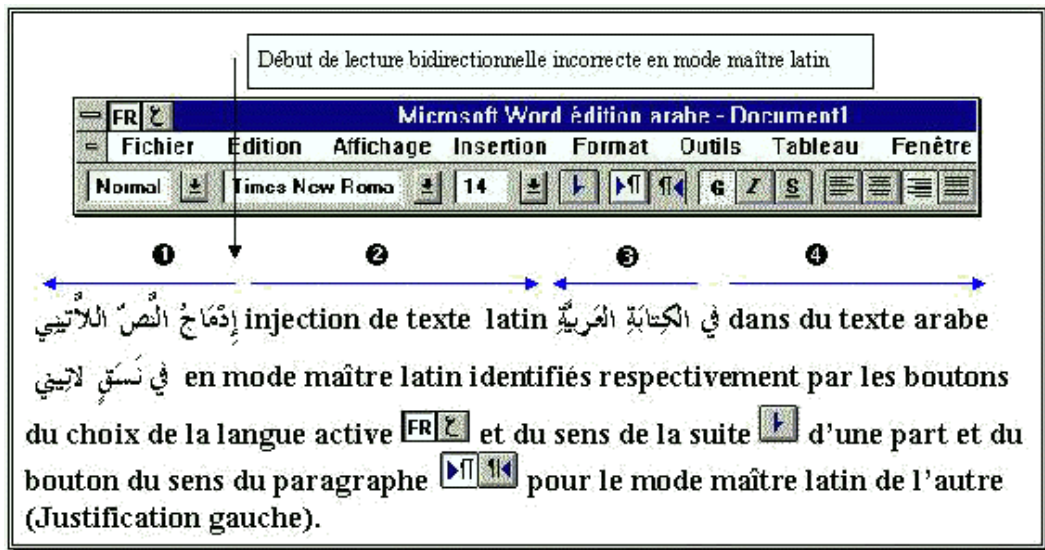


Figure 151 : Incise arabe incorrecte en mode maître latin

- Un mode maître arabe de gauche à droite avec insertion en latin de droite à gauche : ceci constitue un exemple opposé au précédent pour démontrer l'incohérence d'un mode maître d'édition qui commence par un texte dans la langue injectée.

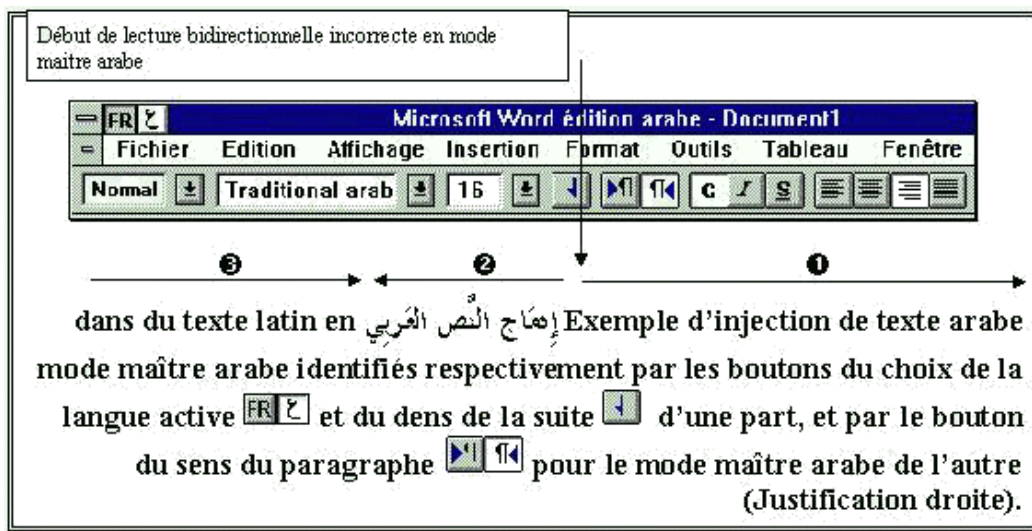


Figure 152 : Incise latine incorrecte en mode maître arabe

2.3.3 L'effet de pivotement graphique et sémantique dans un texte Bidi (Swapping)

L'une des conséquences majeures de la bidirectionnalité est l'effet de pivotement observé au niveau du rendu visuel d'un texte Bidi. Le pivotement s'obtient au niveau de la suite des caractères dans un mot. C'est le "swapping" graphique. Il s'obtient aussi au niveau d'une suite de mots. C'est le "Swapping" sémantique.

2.3.3.1 Le Swapping Graphique

Le pivotement graphique est le résultat évident de deux facteurs : les caractères neutres et les paires des caractères symétriques qui accompagnent le texte Bidi..

Par caractères neutres l'on sous-entend les caractères qui n'ont pas une directionnalité propre, et qui s'alignent par conséquent en fonction de la direction globale du texte. Les caractères neutres sont classés en trois catégories distinctes^[29] : les symboles de ponctuation avec le caractère d'espacement et les paires de caractères symétriques, les chiffres et les séparateurs décimaux, et enfin les symboles de traçage de lignes ou de boîtes.

Les signes de ponctuation sont généralement des caractères neutres qui prennent la direction du mode d'édition maître actif. Une exception peut être formulée à l'égard des trois caractères contextuels "Virgule", "Point virgule" et "Point d'interrogation" qui, s'ils prennent la direction du mode maître actif, changent aussi en forme du glyphe d'un contexte à un autre. Leur neutralité n'est donc pas pratiquée car ils se relaient en duos avec leurs équivalents linguistiques en fonction du mode maître actif.

Les chiffres constituent également une exception car s'ils prennent la direction du mode maître actif, leur mécanisme interne reste fidèle à la directionnalité gauche-droite.

Le "Swapping" des caractères neutres et des paires de caractères symétriques est géré différemment par MS DOS et MS Windows.

Arabic MS DOS gère les caractères neutres selon les convenances de l'utilisateur à travers son interface NLS (National Language System) qui permet de gérer entre autres la forme des chiffres, l'invocation des pages de codes, la configuration des imprimantes en modes bilingue.

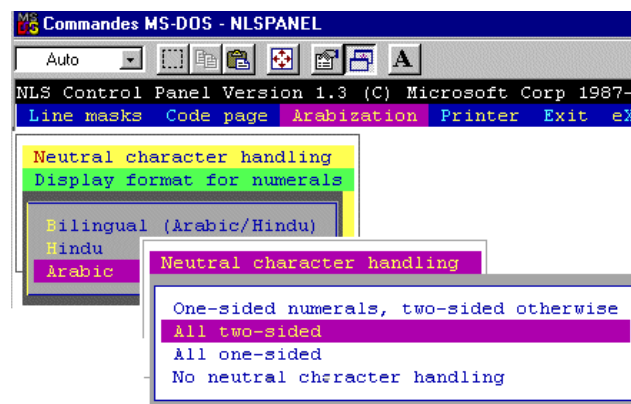


Figure 153 : L'arabisation par NLSPANEL de Arabic MS-DOS

[29] Microsoft MS DOS User's Guide : Arabic Supplement. Microsoft Corporation, 1988. p.18

L'exemple suivant démontre l'effet de pivotement des signes de ponctuation et des paires de caractères symétriques entre les deux modes maîtres sous Arabic MS DOS.

MML		MMA
#2302	1	2302 #
/auteur	2	auteur/
25 قطعة	3	قطعة 25
مختار؛	4	؛ مختار
?question	5	question?
[Et.al]	6	{.Et.al}
{تاريخ}	7	{تاريخ}
{champs}	8	{champs}
{حقل}	9	{حقل}

Tableau 44 : L'effet de Swapping sous Arabic MS DOS

- ❑ Les lignes 4 et 5 montrent l'alternance entre les duos de caractères contextuels de ponctuation d'un mode maître à l'autre;
- ❑ La ligne 5 illustre le cas d'un **pivotement vertical** unique d'un caractère contextuel, en l'occurrence le point d'interrogation. Un pivotement en double est également obtenu par les caractères « virgule » et « point virgule » arabes qui **pivotent verticalement et horizontalement** afin d'apparaître conformément à la contextualité arabe. D'ailleurs, dans la table de code étendue ASMO 708 (équivalente à ISO 8859-6) les caractères spéciaux n'ont pas été reproduits dans la zone GD excepté les caractères particuliers du signe monétaire international (position 10/04) qui substitue le symbole Dollar (position 02/04), le point d'interrogation (position 11/15), caractère contextuel à pivotement vertical par rapport au point d'interrogation latin (position 03/15), la virgule (10/12) et le point virgule (11/11), deux caractères à pivotement horizontal et vertical par rapport aux caractères latin de même fonction (03/11 & 02/12).
- ❑ Les lignes 6 et 7 montrent l'effet de pivotement des paires de caractères symétriques. Le crochet d'ouverture "[" en MML a subi un effet de pivotement en MMA pour garder la même fonction d'ouverture;
- ❑ Les deux points de la ligne 7, qui marquent normalement les abréviations, ont subi l'effet de pivotement en MMA afin de garder leur position de fin de mots;
- ❑ Dans les exemples 8 et 9, le pivotement n'est pas visible bien qu'il a eu lieu, car la paire des caractères symétriques ne change pas de visu dans un swapping horizontal ni vertical.

Les deux colonnes du tableau précédent relèvent les points de détails suivants :

- ❑ La directionnalité gauche-droite est celle du mode maître d'affichage latin (partie gauche) ;
- ❑ La directionnalité droite-gauche est celle du mode maître d'affichage arabe (partie gauche) ;
- ❑ Les caractères spéciaux (# / { } []) changent de position en fonction du mode maître d'affichage sans changer de forme. Ceci est expliqué par le fait que ces caractères détiennent des points de codes uniques dans les tables des codes latin et arabe (Voir

tables [ISO 8859-6=ASMO 708](#)), donc seul leur emplacement est défini par la direction du contexte d'édition actif ;

- Par contre les caractères spéciaux contextuels (, ; ?) changent de position et de forme avec le changement du mode d'édition. Ceci explique l'existence de ces caractères en duos contextuels arabe et latin. L'algorithme de l'analyse contextuelle effectue la transition entre une forme et une autre en fonction du mode maître d'édition.

La gestion des caractères neutres pose un problème de pivotement de direction quand ils sont positionnés au début ou à la fin d'une incise. Si le caractère neutre est au milieu du texte inséré, il prend automatiquement la direction de l'incise dans laquelle il est utilisé. Par contre, s'il est au début ou à la fin de l'incise, il prend la direction du mode maître actif.

Sous MS Windows par contre, les algorithmes de l'analyse contextuelle et du rendu visuel ou imprimé agissent différemment sur cet aspect de pivotement graphique. L'exemple suivant effectué sur MS Excel illustre cette variabilité dans la direction de l'écriture et l'orientation du texte.

E	D	C	B	A	
Context/GD	Context/DG	Justif.Context	Justif.Gauche	Justif.Droite	
1998	1998	1998	1998	1998	1
#2302	2302#	#2302	#2302	#2302	2
/Auteur	Auteur/	/Auteur	/Auteur	/Auteur	3
/مؤلف/	/مؤلف/	/مؤلف/	/مؤلف/	/مؤلف/	4
?Question	Question?	?Question	?Question	?Question	5
{Et.al.}	{.Et.al}	{Et.al.}	{Et.al.}	{Et.al.}	6
{تاريخ}	{تاريخ}	{تاريخ}	{تاريخ}	{تاريخ}	7
{Champs}	{Champs}	{Champs}	{Champs}	{Champs}	8
{حقن}	{حقن}	{حقن}	{حقن}	{حقن}	9
بحرف	بحرف	بحرف	بحرف	بحرف	10
سؤال؟	سؤال؟	سؤال؟	سؤال؟	سؤال؟	11
25 قطعة	25 قطعة	25 قطعة	25 قطعة	25 قطعة	12
25pièces	pièces25	25 pièces	25 pièces	25 pièces	13
					14

Tableau 45 : Effet de Swapping sous MS Excel

- Les colonnes A et B démontrent le déplacement en bloc des chaînes de caractères sans aucun pivotement lorsque la justification est forcée à gauche ou à droite ;
- La colonne C soumet les chaînes de caractères à une justification contextuelle. Seuls les termes commençant par les caractères contextuels arabes "point virgule" et "point d'interrogation" affichent une justification droite. Tous les autres caractères spéciaux, même s'ils précèdent des chaînes de caractères arabes, affichent une justification gauche. Ceci dénote de leur codes binaires puisés dans la partie gauche de la table des codes pour les deux modes arabe et latin.
- La colonne D démontre que lorsque le sens de l'écriture est forcé de droite à gauche, la position des caractères spéciaux passe à droite des termes qu'ils introduisent (i.e. lignes de 8 à 9) ou qu'ils terminent (i.e. ligne 7) ;
- Dans la même colonne D, l'exemple 9/D ne semble pas avoir subi de pivotement car la paire de caractères symétriques n'est pas perceptible dans aucun mouvement de pivot, mais les exemples 7/D et 10/D démontrent le contraire ;

- ❑ Entre les cas 2/D et 15 ou 14/D une analogie est également possible : dans le premier cas le chiffre n'est pas traité comme caractère neutre car le dièse qui le précède (#) l'est aussi. Ce qui fait que entre deux caractères neutres, le premier dans la chaîne des caractères pivote autour de l'autre. Les exemples 14/D et 15/D illustrent cette priorité : les chiffres dans ces cas précis pivotent autour des lettres qui sont purement contextuelles.
- ❑ La colonne E présente cependant une anomalie à ces règles d'affichage : les caractères spéciaux contextuels arabes ne subissent pas le même effet de l'algorithme Bidi qui les exclut du pivotement bidirectionnel. Les exemples des lignes 11 et 12 sont restés toujours constants.
- ❑ Une autre anomalie (par contre justifiée cette fois) est observée également entre les lignes 1, 13 et 14 : la position et la direction des chiffres.
- ❑ À part le fait que le chiffre de la ligne 1 soit trié en priorité (nous en avons expliqué la raison sous la rubrique relative au tri), le fait qu'il soit aligné à droite malgré le changement du sens de l'écriture (Colonne C) et malgré sa propre direction interne (de gauche à droite), dénote d'un détail très important pour notre étude : les chiffres absolus (non suivis de caractères alphabétiques) sont automatiquement justifiés à droite dans une feuille de calcul ou dans un champ numérique d'une table de bases de données (ce n'est pas le cas de MS Word!).

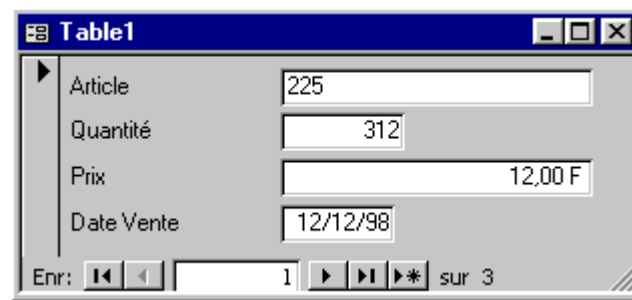


Figure 154 : L'effet de l'identification "Numérique" ou "Texte" des chiffres sous MS Access

Ceci rejoint le point d'analyse que nous allons développer dans un rubrique suivante et qui met en exergue l'évolution de calculabilité de droite à gauche dans les algorithmes algébriques.

2.3.3.2 Le Swapping sémantique

Le swapping sémantique est l'inversion dans l'ordre des phrases qui résulte d'un changement de direction globale d'un texte Bidi. Cette situation est très courante au moment de l'insertion d'incises. Sa complexité est issue de deux phénomènes essentiels : les bris au niveau des incises et l'unicité directionnelle des chiffres. Ces deux aspects constituent deux niveaux de difficultés dans le dialogue avec l'interface machine.

Les bris de texte :

Comme nous l'avons vu en abordant le rapport entre stockage logique et rendu physique des caractères, les bris d'incises constitue une difficulté majeure dans l'affichage des textes multilingues. Deux directions opposées, deux modes d'édition opposés, deux modes de justification opposés... sont d'ordre à rendre le traitement multilingue simultané quelque peu difficile.

Les deux exemples suivants démontrent le mécanisme de bris (simples et complexe) au niveau d'une incise. Le bris d'un texte uniforme ne constitue aucune difficulté car, selon le mode maître actif, le bris va forcer un retour à la ligne dans le sens global du document. Par contre le bris d'une incise d'une directionnalité opposée posera le problème de l'emplacement et du sens de la continuité du texte coupé sur la ligne suivante.

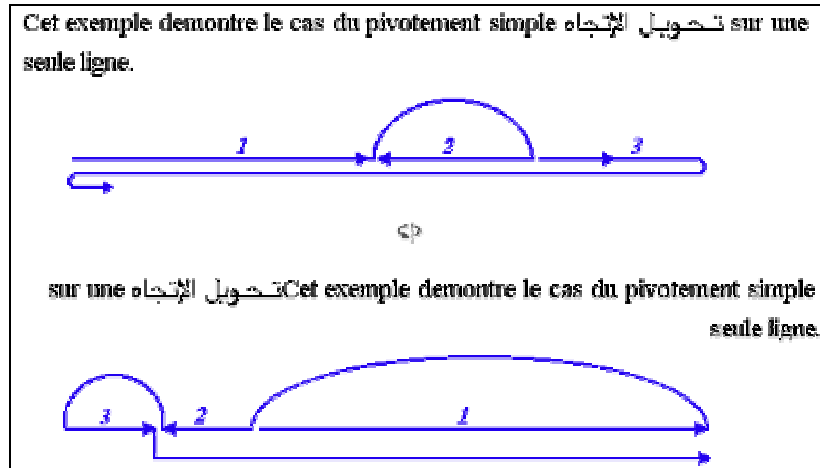


Figure 155 : Pivotement d'une incise sur la même ligne

Par contre, dans les deux exemples suivants, le pivotement a lieu sur une incise avec un bris sur deux lignes. Dans le premier cas, l'agencement est correct car le début du texte de chaque langue commence à l'emplacement qui convient à sa lecture normale (début de ligne) et son flux évolue conformément à la directionnalité normale de chaque langue.



Figure 156 : Bris normal d'incise arabe en MML

Dans le deuxième exemple, bien que le flux du texte et le retour à la ligne soient conformes à une directionnalité normale, les débuts du texte hôte et de l'incise sont opposés à leurs points de départs normaux. Le texte Bidi ne correspond plus à une juxtaposition faciale des deux chaînes linguistiques car les deux commencements sont au milieu de la ligne. Sur le plan sémantique, le sens de la phrase devient alors moins perceptible.

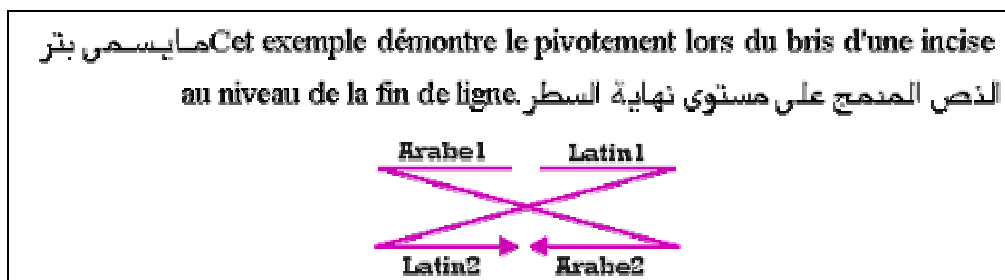


Figure 157 : Pivot anormal d'une incise arabe en MML

Les bris de chiffres :

La complexité des deux derniers cas demeure relative devant les bris des chiffres dans une incise arabe. Cette particularité de l'incise arabe chiffrée est issue de la juxtaposition entre la direction droite-gauche de l'incise et la direction unique gauche-droite des chiffres.

Dans l'exemple suivant, le bris au niveau des chiffres dans l'incise arabe a engendré un pivotement successif des blocs de chiffres isolés par rapport à la suite contiguë de départ.

L'exemple « A » illustre une incise arabe avec un bloc chiffré dans un mode maître latin qui inclut la même série de chiffres. La forme du curseur qui annonce la direction de la suite au niveau des chiffres relève que le mode d'affichage actif est un mode visuel.

Dans l'exemple « B », un espacement est introduit entre les trois derniers chiffres du nombre en mode latin. Le résultat est un décalage vers la droite des trois chiffres. La directionnalité gauche-droite des chiffres et des lettres reste ainsi intacte.

La même opération est effectuée sur le même nombre en mode arabe. Le résultat est un pivotement des trois chiffres isolés vers la gauche du bloc restant des chiffres.

L'explication de ce phénomène est l'insertion d'un caractère espace entre les chiffres. L'espace est un caractère neutre qui, une fois inséré, divise la chaîne des chiffres en deux blocs non contigus. En mode latin, les deux blocs résultants se suivent dans une orientation gauche-droite normale. En mode arabe, les deux blocs de chiffres séparés par l'espace inséré sont invertis pour respecter la directionnalité droite-gauche de l'incise arabe tout en gardant leur structure interne d'une directionnalité standard de gauche à droite.

L'exemple « C » présente une segmentation en plus du bloc de chiffres. Si la linéarité est sauvegardée en mode latin, le mode arabe force une inversion supplémentaire pour rejeter le nouveau bloc de chiffres à gauche de la suite chiffrée.

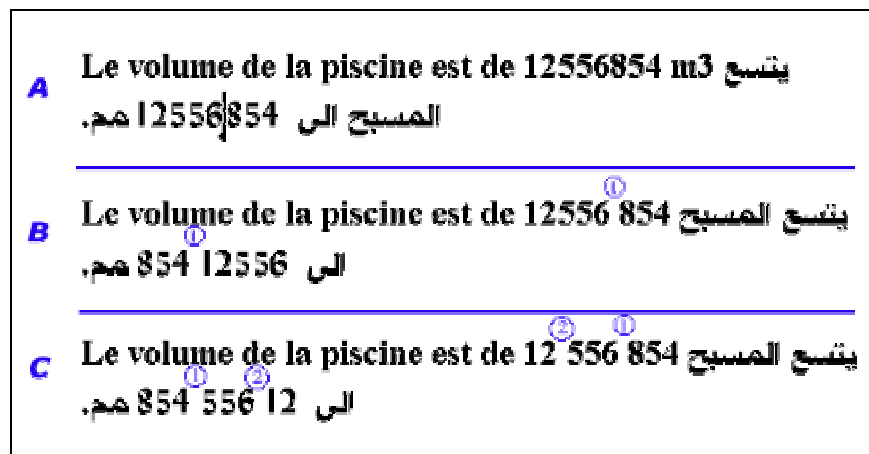


Figure 158 : Le pivotement des chiffres dans une suite arabe

Cet aspect de pivotement des chiffres arabes est d'ordre à créer des perturbations dans la structure sémantiques des textes multilingues. L'exemple d'une fonction « Rechercher-Remplacer » effectuée pour insérer un espace entre l'indicatif régional et les numéros de téléphone engendrera un revirement dans la structure des numéros téléphoniques. La possibilité

d'aérer un chiffre long en espaçant les centaines des milliers et les milliers des millions devient alors impossible car la valeur du chiffre se trouvera ainsi complètement faussée.

Sans caractères neutres	Avec caractères neutres
Tél : 01234567 الهاتف : 01234567	Tél : (01)234567 الهاتف : 234567(01)
Prix : 98765432 D. السعر : 98765432 د.	Prix : 98 765 432 D. السعر : 432 765 98 د.

Tableau 46 : Effet de swapping sur les chiffres en MMA

Ces derniers exemples sont en réalité le prélude d'une partie entière que nous réservons aux chiffres et leurs méthodes de traitement au niveau logique interne et visuel externe au niveau de l'interface Homme-Machine. La partie suivante sera réservée à cet aspect d'usage des chiffres dans leur relation intrinsèque avec les phénomènes de l'information et de la communication.

3. LES CHIFFRES, LA BIDIRECTIONNALITÉ ET L'INTERNATIONALISATION : LA CONFLICTIVE CULTURELLE

Ce n'est pas pour rappeler les définitions standards de l'usage des chiffres dans les formules algébriques que nous évoquerons les valeurs et les attributs des uns ou des autres. Notre objectif est de faire valoir une hypothèse dont le fondement stipule une reconsidération des rapports entre un contenant conventionnel et un contenu conjoncturel. En d'autres termes, nous développerons un questionnement sur les formes conventionnelles d'usage des chiffres dans la quasi-totalité des systèmes informatiques actuels (traitement logique et visuel de gauche à droite) et la nature des rapports qu'elles ont avec la nature conjoncturelle et culturelle de leur usage. Entre le conventionnel et le conjoncturel, nous observons déjà une série d'anomalies que nous rapportons comme levée du rideau sur un des aspects souvent ignorés dans la définition et l'évaluation des interfaces Homme-Machine arabe/latines.

Si nous réservons une partie à part à ce phénomène de chiffres dans une optique de traitement bidirectionnel, c'est pour mettre en exergue un phénomène techno-culturel que nous avons jugé intéressant pour soulever des controverses et des polémiques de conception et d'usage d'interfaces issues du principe de l'unidirectionnalité gauche-droite des chiffres dans tous les modes linguistiques possibles. Nous nous contenterons dans ce chapitre de soulever les anomalies et les incohérences culturelles et techniques de l'usage des chiffres dans un contexte multilingue et surtout bidirectionnel qui pourraient faire objet de recherches plus approfondies dans des disciplines relevant des domaines des sciences humaines (histoire, sociolinguistique...). Nous partons de notre profil d'utilisateurs avertis des nouvelles technologies de l'information et de notre objectif primordial issu du domaine de l'information et de la communication pour évaluer les « pour » et les « contres » d'un domaine qui nous est en quelque sorte tributaire dans une partie de son champ d'action : l'évaluation des interfaces Homme-Machine.

Étant donnée l'importance stratégique des chiffres et de leur traitement dans un contexte informatique, nous reprendrons l'histoire des chiffres depuis leur origine pour mettre l'accent sur deux aspects fondamentaux de leur usage : d'une part le fondement des systèmes numériques et

la forme de représentation des chiffres et d'autre part le mécanisme de la directionnalité algébrique et alphanumérique.

3.1 LES SYSTEMES NUMERIQUES ET LA REPRESENTATION GRAPHIQUE

Il est évident que l'acte de compter a surgi dans la pratique des hommes bien avant l'écriture. Par la nature même de son vécu quotidien, le fait de quantifier a toujours accompagné l'homme dans la découverte de son environnement matériel. À part le déchiffrement sensoriel des objets qui l'entourent, la faculté cognitive de l'homme fait fonctionner un mécanisme de recensement implicite. La complexité progressive de la société a été derrière l'invention des mécanismes artificiels de comptage et de recensement que l'homme a développé selon des méthodes et des logiques diverses pour instaurer des théories entières de mesure qui se basent sur l'arithmétique quantitative et ses méthodes de représentation adéquates. La civilisation humaine est féconde d'exemples et de modèles dont les effets même partiels et secondaires sont détectables dans les systèmes numériques et les mécanismes de calculs des mesures modernes.

3.1.1 L'histoire des systèmes numériques

Les systèmes numériques ont généralement servi pour la quantification des valeurs comptables. Un système numérique est défini selon sa base. La base d'un système numérique détermine le nombre de symboles différents qu'il utilise. Les systèmes numériques les plus primitifs ont été découverts dans les tribus australiennes qui utilisaient des systèmes généralement à base 2 et exceptionnellement à base 3. Ce qui implique que ces systèmes utilisaient 2 ou 3 symboles pour représenter les chiffres individuels. Toute valeur différente de 1, 2 ou 3 est exprimée en tant que multiple ou répétition de ces deux ou trois symboles.

Avant d'arriver à l'état actuel des systèmes numériques contemporains les plus utilisés (décimal et binaire), l'histoire recense un grand nombre de systèmes numériques à bases multiples qui varient entre 2, 3, 4, 5, 10, 12, 20 et 60. Selon l'étude accomplie par W.C. Eels, cité par Shani Welch^[30], sur les sociétés américaines primitives, parmi les 307 systèmes numériques identifiés, 146 étaient à base 10 ; 106 étaient à base 5, 20 ou des combinaisons entre les bases 5, 10 et 20. Cette particularité des systèmes numériques sur des bases en multiples de 5 est expliquée par le phénomène du comptage sur les doigts de la main que pratiquaient les hommes avant l'invention des outils de comptage sophistiqués.

Le développement des sociétés primitives et les besoins grandissant de recensement et de notations numériques complexes a fait apparaître de nouveaux mécanismes allant du principe de regroupement simple au système des valeurs positionnelles. Nous décrivons quelques cas ci-après.

3.1.1.1 Le compte par itération (boucles)

Les Égyptiens ont développé l'un des premiers systèmes numériques qui constitue l'une des grandes phases de développement dans l'usage des chiffres. Le système **hiéroglyphique** est basé sur la répétition des symboles 1, 10, 100, 1000, 10 000, 100 000 et 1 000 000 dans une structure combinatoire orientée de droite à gauche et sans la présence d'une valeur nulle.

[30] Shani WELCH. History of numeral systems. 1995.

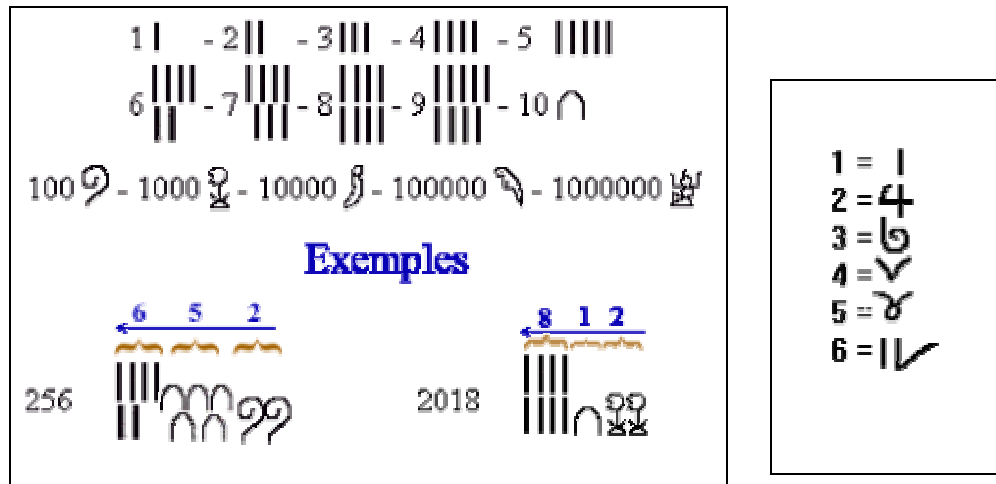


Figure 159 : Les systèmes numériques hiéroglyphique et hiératique égyptiens

Adaptées de : Social and historical aspects of mathematics. <http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>

Quand ce système de transcription hiéroglyphique s'est développé pour prendre les formes de l'écriture hiératique, les Égyptiens ont fait évoluer leur système numérique à base de répétition vers le système de chiffrement dans lequel chaque valeur numérique est codée par un symbole particulier. Les formes hiératiques s'étendent jusqu'à une valeur de 10^3 (Dix au facteur trois).

L'influence de ce modèle sur le système des chiffres chez les romains est très notoire. À part l'inversion de la directionnalité vers le sens gauche-droite, les mêmes critères de répétition de valeurs numérique par tranches répétitives (entre les valeurs 1, 5, 10, 50, 100, 500 et 1000 dans ce cas) et l'absence de la valeur nulle sont reproduits. L'inconvénient du système romain c'est qu'il était beaucoup trop rudimentaire pour faire les vraisemblables développements subséquents dans la science et la technologie.

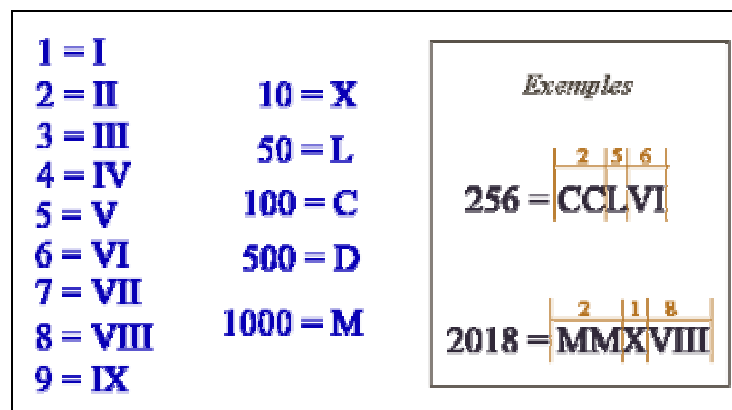


Figure 160 : Le système numérique romain

En effet, les chiffres romains ainsi que la plupart des autres systèmes numériques d'antan ne s'accordent pas avec les mécanismes du calcul écrit. Ceci est du à la nature statique de leurs chiffres de base qui ne sont en réalité que des enregistrements abrégés des résultats des calculs obtenus au moyen d'un abaque ou d'un tableau de compte.

3.1.1.2 Le compte par paires

Ce système se base sur la subdivision par des ensembles de 2. Un nombre trois serait identifié comme un ensemble 2 et un sous-ensemble 1.

Bien que complètement dépassé par les nouveaux systèmes de compte bien plus sophistiqué, ce système de comptage est encore utilisé dans certaines tribus africaines, australiennes et sud-américaines selon la structure suivante :

Un (1),
 Deux (2),
 Deux - Un (3)
 Deux - Deux (4)
 Deux - Deux - Un (5) etc.

3.1.1.3 Le néo-compte à base de paires

Dans un système à base de paires la valeur 6 serait égale à $2 + 2 + 2$. La valeur 8 serait égale à $2 + 2 + 2 + 2$. Certains systèmes à paires ont évolué pour traiter des valeurs de groupes multiples de 2. Ainsi la valeur 6 devient 2 fois 3 et 8 devient 2 fois 4.

L'une des explications données à cette évolution est l'usage des bâtonnets que les Égyptiens ont utilisés depuis 3000 ans A.C. Ce qui donne à une représentation des valeurs 6, 7 et 8 les formes suivantes :

111	1111	1111
111	111	1111
6	7	8

Figure 161 : Représentation des valeurs numériques par bâtonnets

3.1.1.4 Le compte à base 4 et à base 12

Le compte à paires était très limitatif. D'autres systèmes étaient apparus sur des base 4 et 12. Dans ces systèmes, les nombres sont traités comme suit :

1
 2
 3
 4
 $5 = 4 + 1$
 $6 = 4 + 2$
 $7 = 4 + 3$
 $8 = 4 \times 2$
 $9 = 4 \times 2 + 1$
 $10 = 4 \times 2 + 2$
 $11 = 4 \times 2 + 3$
 12
 $13 = 12 + 1$
 $14 = 12 + 2$

3.1.1.5 Les systèmes de compte 5-10 et 5-20

Le système 5 - 10 identifie par termes indépendants les valeurs entre 1 et 5. Les nombres de 6 à 9 sont identifiés par des mots composés des 5 premiers éléments. Les termes désignant le chiffre 10 et ses multiples sont composés comme suit :

Le système de compte 5 - 20 prévoit 5 principes de comptage jusqu'à 19 :

Les Aztèques comptent 1, 2, 3, 4, 5, 5 + 1, 5 + 2, 5 + 3, 5 + 4, 10, 10 + 2, 10 + 3 ... 20, 20 + 1, 1 x 20 + 10 ... 2 x 20 + 10.

Ce système peut être représenté par un schéma de boîtes en escalier :

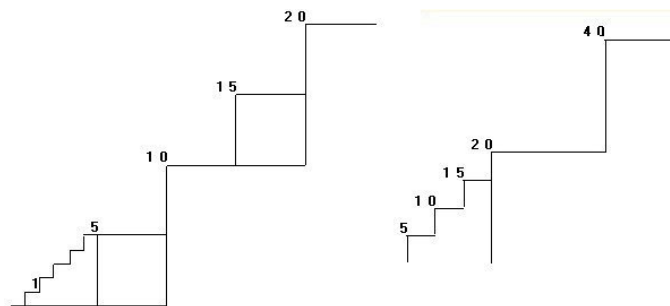


Figure 162 : Systèmes de calcul 5-10 & 5-20

Source : Social and historical aspects of mathematics. <http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>

Les Babyloniens utilisaient le système *sexagésimal* (basé sur le nombre 60). Les romains utilisaient parfois le système *duodécimal* (basé sur le nombre 12). Les mayas utilisaient le système *vigesimal* (basé sur le nombre 20). Le système binaire (basé sur le nombre 2) est utilisé aujourd'hui dans les systèmes informatiques. Plus la base du système est grande plus le nombre de symboles augmente mais moins de signes y sont nécessaires pour représenter une valeur quelconque.

3.1.1.6 Le système 10-60

Utilisé par les Sumériens et les Babyloniens il y a déjà 5000 ans, le système de compte 10-60 est à sa base un système décimal. Seulement la valeur 60 y représente une nouvelle unité clé.

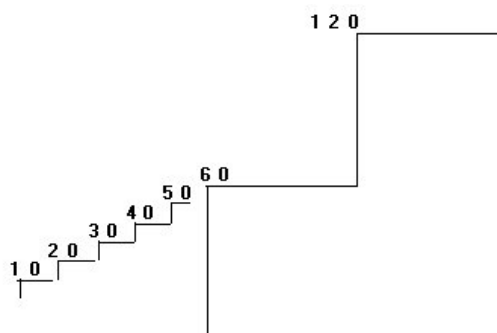


Figure 163 : Le système de compte 10-60

Source : Social and historical aspects of mathematics. <http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>

Les traces de ce système ainsi de certains systèmes de compte à base 20 sont encore perceptibles dans les langues d'Europe. En français la structure de comptage change à partir de la valeur 60. Le chiffre 70 (Soixante-dix) est segmenté en soixante plus dix ; la valeur 80 (quatre-vingts) est décomptée 4 fois 20 ; le nombre 90 (quatre-vingt-dix) est structurée en mode combinatoire de 4 fois 20 plus 10. Les anglais utilisent aussi l'expression "3 score" pour dire 60 et les danois pratiquent un comptage par vingtaine à l'instar des basques, irlandais et écossais^[31]. D'autre part, si nos montres affichent 12 heures par jour par des sous unités temporelles de 60 minutes et de 60 secondes, et si les mesures angulaires sont exprimées en degrés et secondes par multiples de 60 (180°, 360°), l'origine des ces valeurs de mesure ne peut être qu'une inspiration d'un système ancien, en l'occurrence celui des babyloniens.

3.1.1.7 Le système décimal : les valeurs positionnelles

On attribue souvent le système décimal au nombre 10 des doigts des deux mains qui représentent le premier "ustensile" humain de comptage.

Le principe du système décimal est le fait que la valeur 10 soit considérée comme une nouvelle unité à partir de laquelle le comptage devient cyclique. La valeur 100 (10 fois la valeur 10) est une autre unité de comptage cyclique. Les multiples de 10 sont ainsi traités de la même manière que les valeurs de 1 à 9.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
'	'	'	'	'	'	'	'	'	
91	92	93	94	95	96	97	98	99	100

Tableau 47 : Le système décimal à base 10

La position d'un symbole dans un nombre à base 10 détermine la valeur de ce symbole en terme de la puissance de sa base. Dans ce système le symbole Zéro joue un double rôle. Il représente la valeur nulle et il sert aussi à désigner les multiples de la base 10 : 100, 1000, 10 000 etc.

Le système numérique décimal utilise 10 chiffres de 0 à 9. Les nombres au-delà de 9 sont en réalité des combinaisons des même chiffres. Le nombre 10 est la combinaison de 1 et de 0 dans un ordre de position stricte. Le 1 est situé en deuxième position à partir de la gauche. Cette position lui attribue la valeur 10 selon la structure des valeurs positionnelles Unité, Dizaine, Centaine, Millier etc.

La valeur des emplacements dans la structure décimale augmente par un facteur de « 10 » à partir de la position la moins importante située à droite. En d'autres termes, une valeur dans une position quelconque est multipliée par le facteur « 10 » si elle passe à la position suivante à sa gauche. Dans l'exemple 4 du tableau suivant, la valeur du symbole décimal « 5 » sous la colonne du poids « 100,000 » est le multiple par le facteur « 10 » de la valeur du même symbole décimal « 5 » sous la colonne du poids « 10,000 » à sa droite.

[31] Social and historical aspects of mathematics. University of Wolverhampton. School of Computing and Information Technology. Module MM2217 <http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>

100,000	10,000	1,000	100	10	1	Valeurs positionnelles
0	0	0	0	5	5	Ex.1 55 = 50 + 5
0	0	0	5	5	5	Ex.2 555 = 500 + 50 + 5
0	0	5	5	5	5	Ex.3 5555 = 5 000 + 500 + 50 + 5
5	5	0	0	5	5	Ex.4 550 055 = 500,000 + 50,000 + 0,000 + 000 + 50 + 5

Tableau 48 : Valeurs positionnelles dans un système décimal

La valeur décimale 550,055 de l'exemple 4 est représentée par une suite de symboles pris dans la marge des 10 symboles décimaux entre « 0 » et « 9 ». Si nous attribuons à chaque position sa valeur correspondante, nous obtiendrons l'équation suivante : $550,055 = 5 \times 100,000 + 5 \times 10,000 + 0 \times 1000 + 0 \times 100 + 5 \times 10 + 5 \times 1$.

En définitive, le système numérique décimal est construit sur l'ordre de calcul des valeurs positionnelles à base 10 conformément à la formule suivante :

Poids	Formule	Développement
1	10^0	$10 * 0$
10	10^1	$10 * 1$
100	10^2	$10 * 10$
1,000	10^3	$10 * 10 * 10$
10,000	10^4	$10 * 10 * 10 * 10$
100,000	10^5	$10 * 10 * 10 * 10 * 10$
...	10^n	$10 * n$

Tableau 49 : Mécanisme de calcul du système numérique décimal

Cette structuration à valeurs positionnelles du système numérique décimal s'applique aussi au système numérique binaire.

3.1.1.8 Le système binaire

Le système binaire utilise les deux symboles « 0 » et « 1 » comme marqueurs de positions dans une combinaison binaire ordonnée dans une structure de poids. Tout nombre entier peut être exprimé dans un système binaire sous forme de la somme des différents poids de la valeur 2 ([Tab.50](#)).

Le système binaire est un système à valeur positionnelle. La valeur des emplacements dans la structure binaire augmente par un facteur de « 2 » à partir de la position la moins importante. En d'autres termes, une valeur dans une position quelconque est multipliée par le facteur « 2 » si elle passe à la position suivante à sa gauche. Dans l'exemple 4 du tableau suivant, la valeur du symbole binaire « 1 » sous la colonne du poids « 32 » est le double de la valeur du même symbole binaire « 1 » sous la colonne du poids « 16 » à sa droite.

256	128	64	32	16	8	4	2	1	Valeur positionnelle
0	0	0	1	1	0	1	0	0	Ex.1 52 = 32 + 16 + 4
0	0	0	1	1	0	1	0	1	Ex.2 53 = 32 + 16 + 4 + 1
0	0	0	1	1	0	1	1	0	Ex.3 54 = 32 + 16 + 4 + 2
0	0	0	1	1	0	1	1	1	Ex.4 55 = 32 + 16 + 4 + 2 + 1

Tableau 50 : Valeurs positionnelles dans un système binaire

Dans le même tableau, la valeur de la position suivante à 256 n'est autre que la valeur 512 en tant que multiple de 256 par le facteur « 2 ».

La valeur décimale 55 de l'exemple 4 est représentée en binaire sous forme de « 000110111 ». Si nous attribuons à chaque position sa valeur correspondante, nous obtiendrons l'équation suivante : $55 = 0 \times 256 + 0 \times 128 + 0 \times 64 + 1 \times 32 + 1 \times 16 + 0 \times 8 + 1 \times 4 + 1 \times 2 + 1 \times 1$ (Les poids 256, 128, 64 et 8 étant nuls, ils ne jouissent d'aucune valeur positive). En définitive, le système numérique binaire est bâti sur l'ordre de calcul des valeurs positionnelles à base 2 conformément à la formule suivante :

Poids	Formule	Développement
1	2^0	$2 * 0 (=1)$
2	2^1	$2 * 1$
4	2^2	$2 * 2$
6	2^3	$2 * 2 * 2$
8	2^4	$2 * 2 * 2 * 2$
16	2^5	$2 * 2 * 2 * 2 * 2$
32	2^6	$2 * 2 * 2 * 2 * 2 * 2$
64	2^7	$2 * 2 * 2 * 2 * 2 * 2 * 2$
128	2^8	$2 * 2 * 2 * 2 * 2 * 2 * 2 * 2$
...	2^n	$2 * n$

Tableau 51 : Mécanisme de calcul du système numérique binaire

3.1.1.9 La formule mathématique

Pour résumer le mécanisme des systèmes numériques nous pouvons avancer la formulation mathématique suivante :

Quelle que soit « b » la base d'un système numérique, et « a(j) » représente « n » chiffres de tout nombre de sorte que « j » est prise entre « 1 » et « n », un système numérique est représenté par la formule suivante :

$$\sum (a(j) * b^{(j-1)})$$

Où

$$\sum > j \text{ et } 0 \leq a(j) \leq (b-1)$$

Ce qui signifie que les chiffres valables pour un système numérique quelconque doivent être entre 0 et (b-1) inclusif.

3.1.2. La graphie des chiffres : la confluence des cultures

Pour rester dans l'optique de notre problématique des chiffres, l'intérêt particulier que nous réservons au système numérique décimal et la généralisation de son usage comme alternative

universelle pour toutes les solutions informatiques mono et multilingues, nous pousse à remonter le cours du temps pour revenir à ses origines historiques afin d'y retrouver des traces techniques et culturelles qui pourraient nous faire comprendre ses différentes particularités, ses déboires et ses soubassements civilisationnels et linguistiques qui l'ont marqués au fil du temps. Notre objectif est de situer les origines de certaines anomalies de ce système dans une plate-forme multilingue que nous nous sommes proposées d'étudier en rapport avec l'ergonomie des interfaces Homme-Machines arabe-latines. Le principe de représentation visuelle est le premier point à débattre pour expliquer les origines et la dualité très enracinée dans le monde arabe de l'usage des deux formes de glyphes pour représenter les chiffres : les chiffres hindous et les chiffres arabes.

3.1.2.1 La transcription graphique des chiffres

La transcription graphique des chiffres constitue l'une des controverses internes chez la communauté arabophone. Cette diversité influe considérablement sur les systèmes multilingues arabes latins. La conception elle-même des solutions informatiques multilingues arabe/latines connaît un degré de complexité supplémentaire à cause de la dualité des glyphes des chiffres utilisés entre « Maghreb » et « Machrek ».

En fait, c'est vers le VIII^e Siècle de notre ère, qu'un nouveau style d'écriture des chiffres est apparu chez les Arabes sous influence grecque. Au même moment un deuxième style d'écriture des chiffres, une variante des chiffres brahmanes, est apparu sous le nom générique de "chiffres arabes de l'Est" avec un symbole particulier au Zéro sous forme de point. C'est la forme encore d'usage dans grand nombre de pays arabes et plus particulièrement depuis la partie Est de la Libye jusqu'aux limites orientales de la zone d'usage de l'alphabet arabe.

0	1	2	3	4	5	6	7	8	9	10
•	١	٢	٣	٤	٥	٦	٧	٨	٩	١٠

Figure 164 : Correspondance de chiffres entre systèmes Arabe de l'Ouest et Arabe de l'Est (Hindou)

C'était une version modifiée des chiffres brahmanes. À l'Ouest de l'Arabie, les chiffres « Gobar » dits chiffres arabe de l'Ouest étaient contemporains et de même descendance hindoue que les chiffres arabes de l'Est. Comme l'illustre la figure précédente (Fig.164), le symbole « Zéro » est apparu avec les Hindous sous forme d'un point retransmis en Arabie de l'Est d'où il fut introduit dans l'écriture arabe de l'Ouest. *Encarta Concise Encyclopedia* mentionne que les Mayas utilisaient vers le siècle premier de notre ère une forme ovale avec un arc de cercle dedans comme marque d'une position vide. Cinq siècles plus tard, les Hindous utilisaient un cercle ou un point pour représenter le « Zéro » comme valeur positionnelle nulle.

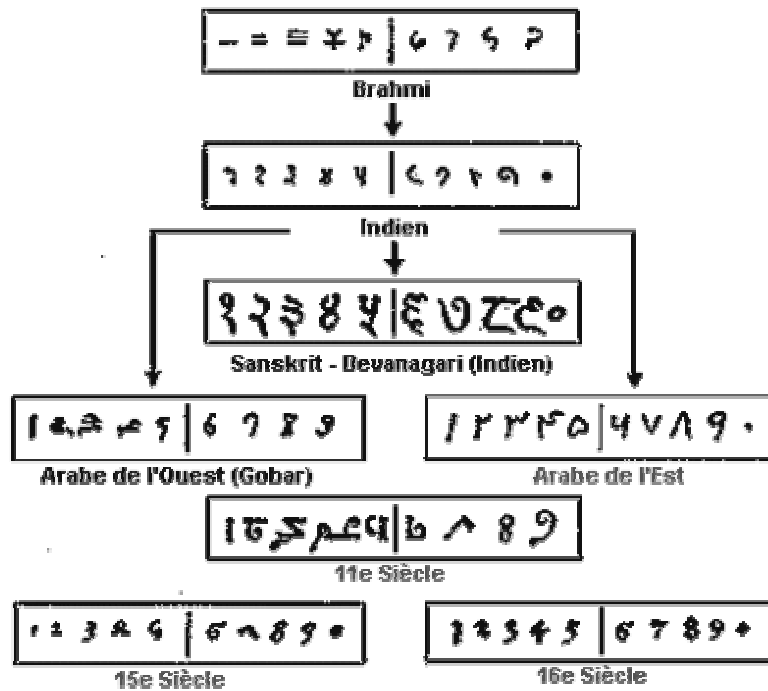


Figure 165 : Evolution historique des chiffres arabes entre 250 A.J. et le 16e Siècle

Adaptée de : Social and historical aspects of mathematics. <http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>

La littérature évoque parfois, sans preuves concrètes à notre connaissance, d'une origine géométrique de la forme des chiffres arabes de l'Ouest inspirée du nombre d'angles qui constituent le chiffre en question. Ainsi le symbole « Zéro » dans sa forme circulaire émane de l'absence d'angles qui le constituent.

0 1 2 3 4 5 6 7 8 9

Figure 166 : La forme des chiffres arabes selon le nombre de leurs angles

3.1.2.2 Sur la route de l'Occident

Nous commencerons à partir du déclin de Rome, quand l'Inde était devenue entre le premier et le cinquième siècle après Jésus-Christ le grand centre de la recherche mathématique. La civilisation hindoue connaissait un développement culturel et une situation commerçante florissante qui allait avoir des répercussions très profondes sur les cultures ultérieures. La contribution la plus importante des Hindous était le développement du système numérique décimal basé sur le principe des valeurs de positions. Seuls les Hindous dans le contexte des civilisations indo-européennes avaient utilisé le Zéro comme valeur nulle indispensable dans toutes les facettes du calcul. L'influence Hindoue sur les systèmes numériques et les théories algébriques allait ensuite largement contribuer à la mise en place des systèmes numériques actuels. Depuis la deuxième moitié du premier millénaire avant notre ère, les tribus indo-européennes se déplaçaient vers le

Nord Ouest de l'Inde portant avec elles une nouvelle langue : le Sanskrit^[32]. En Sanskrit, (la langue savante des Hindous), des nombres-mots (nombres exprimés en mots) désignent les valeurs de 1 à 9, 10, 100 et les multiples de 10 jusqu'à l'ordre de 10^{10} (c'est le poids le plus élevé de l'abaque). En Sanskrit, le mot pour désigner le zéro est « sunya » signifiant « nul » « et il y a peu de doute que le concept du zéro provienne comme le symbole écrit pour la colonne vide de l'abaque. Pour fournir un moyen facile d'accumuler des produits progressifs de multiplication autour du monde, l'usage de l'abaque en déplaçant ces produits supplémentaires toujours orientés vers la gauche, colonne par colonne, l'opérateur remplissait un par un les espaces de perles disponibles et déplaçait l'excès de dix sur les colonnes successives de droite-à-gauche »^[33].

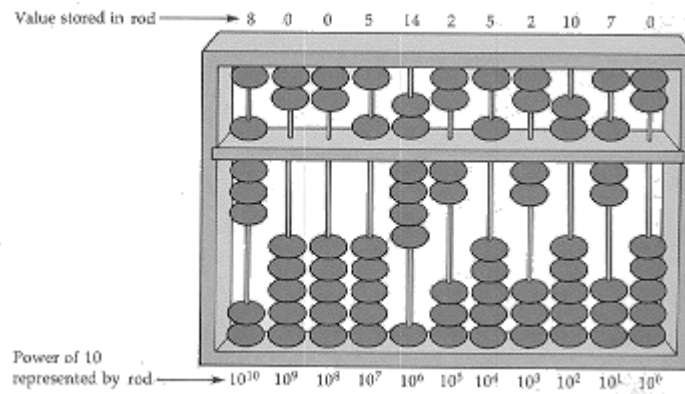


Figure 167 : L'abaque

Dans la même région, des inscriptions Brahmanes trouvées dans le sous continent indien révèlent un système numérique décimal qui existait depuis le troisième millénaire avant notre époque et dont la structure est répartie selon le modèle de valeurs positionnelles. Les savants Brahmanes, les Pythagoriciens de l'Orient, ont fait progresser et perfectionner le système presque jusqu'à sa formulation actuelle. En utilisant les dix symboles numériques de 0 à 9, la seule signification et fonction qu'ils réussissent à développer parfaitement est la formulation de nombres infiniment grands et la résolution d'opérations numériques complexes de manières beaucoup plus simples.

Dans sa progression vers l'Ouest, la méthode des Hindous dans la transcription des chiffres est introduite en Perse vers l'an 750 de notre époque où il subit un mouvement de traduction massive vers la langue arabe. Ainsi, le système des chiffres créé par les Hindous vers l'an 600 de notre époque parvenait jusqu'en Arabie entre le VII^e et le VIII^e siècles où il subit une nouvelle mise au point qui lui donna la forme définitive connue jusqu'à notre époque.

Al Khawarizmi était le premier arabe à avoir expliqué le système numérique hindou dans son livre « *Kitab al-jabr wa al-muqabala* » (§ II-4.1). Ce livre d'arithmétique serait ainsi le premier

[32] « Utilisée, analysée et conservée par une classe de lettrés professionnels appelés pandits, elle a eu ses structures définitivement fixées aux environs de l'ère chrétienne. Elle a été dès lors appelée « sanskrit » qui signifie « construit [selon les règles de la grammaire] ». Elle a été utilisée sans interruption jusqu'à nos jours comme le principal véhicule de toutes les activités intellectuelles : grammaire, logique, exégèse, sciences, techniques, droit, rituel, philosophie, etc., et belles-lettres. Elle a été la langue de relations entre les différentes régions de l'Inde et même de relations internationales en Asie, vers la Chine par l'Asie centrale et dans tout le Sud-Est asiatique. La longévité du sanskrit, l'immensité de son aire d'utilisation, le monopole dont il a joui dans l'expression de tous les genres du savoir humain, la diversité des contacts et des influences qu'il a pu avoir et recevoir font de lui une langue au vocabulaire exceptionnellement riche et de sa littérature la plus abondante et diversifiée du monde ». Extrait de : Pierre-Sylvain FILLIOZAT. Sanskrit : langue et littérature. Encyclopædia Universalis, France S.A, 1995. 20-571 C.

[33] Yasuhiko KIMURA. Néo-tech : le Zéro Philosophique. <http://www.neo-tech.com/french-zero/front.html>

traité arabe à avoir expliqué le système numérique à valeur positionnelle et les opérations qui en dépendent en incluant le « Zéro » comme valeur nulle. Le premier mot du titre de ce traité, « Al-jabr... » est source de désignation de la branche connue des mathématiques en français tel que l'algèbre. Il y démontre clairement comment la valeur d'un nombre change avec le changement de sa position. Son autre traité sur l'arithmétique, « Traité sur le Chiffre », est à son tour le plus récent des travaux arabes connus pour avoir élucidé la numération hindoue de la valeur de la position et les méthodes de calcul.

La route de l'Occident européen pour les chiffres arabes est à plusieurs voies. Certains attribuent le passage des chiffres arabes à l'Occident durant les croisades, d'autres mettent en relief le rôle prépondérant de l'Andalousie Musulmane et son contact avec l'Europe limitrophe. Certains reconnaissent le rôle important de l'Afrique du Nord et ses échanges économiques avec l'Italie proche. Nous pensons que les trois versions se complètent si nous considérons les trois phénomènes historiques suivants :

- ❑ La partie occidentale de l'Arabie était en étroite relation avec les civilisations Omayyade et Abbasside d'où était partie la conquête de l'Ouest à travers l'Afrique du Nord et l'Andalousie. Bien que l'invasion Mauresque de l'Espagne ait introduit les chiffres hindou-arabes et l'algèbre pour la première fois en Europe vers 700 après Jésus-Christ, devançant Al-Khawarizmi d'environ un siècle, il a fallu attendre 400 ans plus tard (vers 1200 après Jésus-Christ.) pour que la traduction latine du "Traité sur le Chiffre" d'Al-Khawarizmi parvienne en Italie du nord et au sud de l'Allemagne par la route de Carthage en Afrique du Nord. C'était en fait suite à la chute de Tolède en 1085, période de la Croisade Occidentale contre les Musulmans en Espagne, que la grande partie des versions arabes de la science grecque aussi bien que des chiffres Hindou-arabes ont été traduits en latin. Il a fallu attendre 200 années supplémentaires pour que la numération hindou-arabe avec le zéro et le concept de la valeur de la position devienne largement diffusé en Europe. L'Afrique du Nord (excepté la Libye) allait subir plus tard dans un mouvement inverse l'influence des andalous qui, après la chute de Grenade en 1492, avaient ramené une science et une culture plus profonde et raffinée avec un système numérique transcrit en chiffres arabes encore d'usage jusqu'à nos jours. C'est l'une des raisons de l'attachement des pays maghrébins aux chiffres arabes de l'Arabie de l'Ouest ;
- ❑ La partie Nord de la péninsule arabe était un croisement de culture et un point de conflits (Croisades) entre le moyen Orient et l'Europe. C'était l'un des points d'infiltration des chiffres de l'Arabie de l'Ouest vers l'Europe à côté des passages andalous et italien. La présence de capitales culturelles et politiques des deux grandes dynasties Omayyade et Abbasside de l'époque (Bagdad et Damas) a également joué un rôle déterminant dans la canalisation d'une culture orientale vers le Nord à travers la Turquie et les Balkans.
- ❑ La partie orientale de l'Arabie développait de plus fortes relations avec les civilisations extrême-orientales persane, hindoue et sino-asiatiques, sources originales des systèmes numériques à valeurs positionnelles dans une transcription graphique restée sous influence brahmane et hindoue. Ceci justifie l'attachement actuel des pays du moyen orient au système numérique oriental hindou.

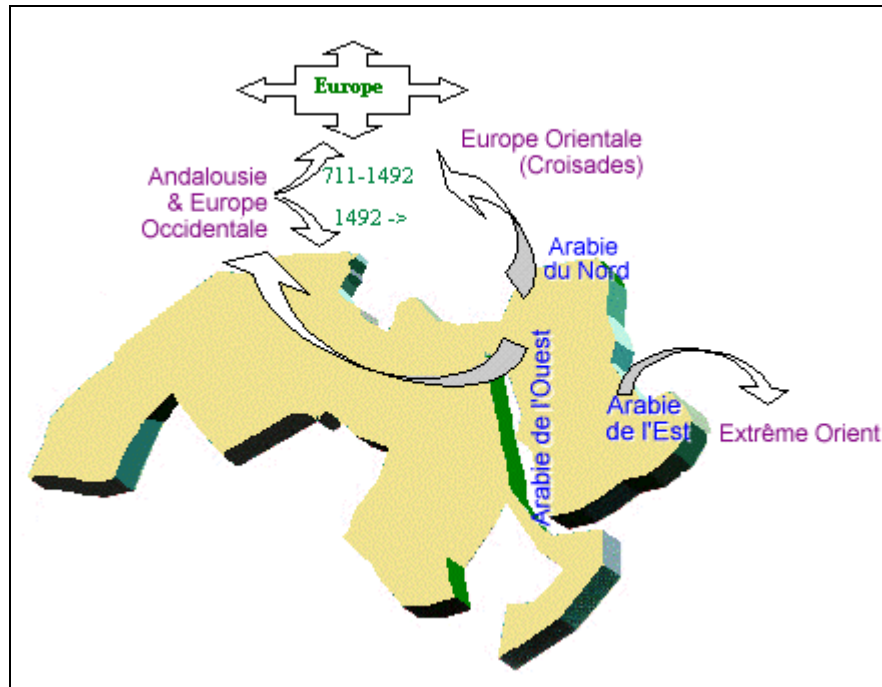


Figure 168 : Migrations des systèmes numériques hindou-arabes

L'histoire rapporte que le plus ancien manuscrit européen avec des inscriptions en chiffres arabes est le "*Codex Vigilanus*" rédigé en Espagne en 976 de notre ère. Selon Struik, cité par Paul J. Gans^[34], le moine et le mathématicien français Gerbert (940-1003)^[35], devenu plus tard le Pape Sylvestre II en 999, les a utilisés de manière très limitée dans certains de ses manuscrits. Le plus ancien manuscrit français d'accès public utilisant les chiffres arabes est daté de l'an 1275.

En Angleterre, le système romain a duré jusqu'à la fin du XVI^e siècle. Ce n'en qu'en 1635 que les chiffres arabes avaient complètement remplacé leurs équivalents romains.

En Italie par contre, malgré la diffusion massive des chiffres romains en Europe occidentale et l'usage du système numérique grec autour de l'adriatique, Léonardo de Pise, connu sous le nom de Fibonacci, a publié son ouvrage "*Liber Abaci*" en 1202 ; un ouvrage d'arithmétique et d'algèbre dans lequel l'influence de Al-Khawarizmi était déterminante.

En définitive, les chiffres arabes ont eu le dernier mot dans le grand conflit qui opposa les adeptes du boulier (ou l'abaque) et les adeptes des algorithmes^[36] pour le maintien ou l'abolition d'un système ou d'un autre. Les facteurs suivants ont tranché pour l'adoption de la nouvelle approche positionnelle des calculs :

- ❑ L'introduction des chiffres « Gobar » de l'Arabie de l'Ouest sur l'abaque^[37] de Gerbert ;
- ❑ La traduction des algorithmes arabes et la rédaction des dissertations sur l'algèbre ;

[34] Paul J.GANS. The Medieval Technology Pages : Arabic Numerals. <http://scholar.chem.nyu.edu/technology.htm>

[35] Né à Aurillac en Auvergne en France, Gerbert régna de 999 à 1003. Il reçut ses premiers cours au Monastère d'Aurillac. il était ensuite pris en charge par un compte espagnol qui l'a inscrit dans des écoles de formation spécialisées à Barcelone, Cordoue et Séville où il apprit les mathématiques et les sciences naturelles des enseignants arabes. *Source* : « The Catholic Encyclopedia ». The Encyclopedia Press, Inc. Version électronique, 1993.

[36] Il s'agit de deux tendances opposées, la première portée sur le maintien du système numérique romain et la deuxième encline à faire usage du nouveau système de chiffres arabes.

[37] Diagramme, graphique donnant par simple lecture la solution approchée d'un problème numérique.

- ❑ L'infiltration au Sud de l'Italie des marchands qui traitaient avec les commerçants de l'Afrique du Nord.

Ce n'est qu'à partir du XIV^e siècle que les chiffres arabes ont connu une extension massive dans toute l'Europe. Cette extension engendrera par moments de rayonnement culturel ou par périodes de décadence un mouvement de retour.

Cet effet de retour (round trip) des systèmes numériques est en fait survenu en Occident et en Orient musulmans. En Occident, la migration des morisques vers l'Afrique du Nord a ancré un système numérique rodé par l'expansion économique et intellectuelle andalouse. En Orient Musulman, à part le phénomène de pèlerinage des populations musulmanes d'Extrême Orient vers la Mecque, la fondation par le deuxième Calife Abbasside El Mansour en 772 après Jésus-Christ de la capitale Bagdad, a permis de créer un grand centre d'échanges commerciaux et intellectuels du monde où l'héritage culturel des nations conquises avait été bien reçu. C'était à Bagdad qu'a commencé l'évolution de la science arabe, assimilant tous les travaux des scientifiques grecs et hindous.

Cet effet de retour aura plus tard un effet déterminant sur la distinction encore en vigueur entre un système oriental à base de chiffres hindous et un système numérique occidental à base de chiffres arabes. La période coloniale du XX^e siècle aura un effet de cristallisation d'un nouveau mécanisme comptable et d'une structuration d'un système numérique arabe hybride. Comme nous le démontrerons dans la partie suivante, ces migrations du système numérique entre l'orient et l'occident se répercuteront sur les systèmes informatiques actuels et influenceront beaucoup le domaine de l'interactivité homme machine dans un contexte multilingue et tout particulièrement le contexte arabe-latin.

3.2 L'UNIDIRECTIONNALITE ALGEBRIQUE ET LA BIDIRECTIONALITE ALPHANUMERIQUE

L'une des particularités très importantes pour notre étude des systèmes numériques est la directionnalité de ce système que les occidentaux qualifient d'inversée, c'est à dire de droite à gauche contrairement au sens de l'écriture des langues européennes de Gauche à Droite. Le chiffre « 1998 » serait écrit dans ce système numérique sous la forme 8991. Les Arabes ont maintenu cette particularité directionnelle par la nature même de la directionnalité Droite-Gauche de leur écriture.

La question que presque personne, selon notre propre investigation, n'a évoquée dans ce genre d'études et à laquelle nous apporterons quelques éléments de réponses est le POURQUOI et le COMMENT le système numérique décimal de notre époque a été reconverti vers une directionnalité Gauche-Droite malgré ses débuts Droite-Gauche. Quelles sont ensuite les conséquences de l'effet de retour que les systèmes numériques d'origine Droite-Gauche ont subi une fois qu'ils avaient été retransmis à leurs sources dans une directionnalité Gauche-Droite ? C'est là l'un des aspects du multilinguisme lourd et de la bidirectionnalité des solutions informatiques actuelles au niveau desquels nous comptons apporter un plus, ne serait-ce que du point de vue de l'approche culturelle et de l'analyse ergonomique, dans l'évaluation des interfaces Homme-Machines arabes-latines.

Nous aborderons cet aspect de bidirectionnalité en ce qui concerne les trois facteurs potentiels de traitement des chiffres : la calculabilité arithmétique et le mécanisme incrémentiel des chiffres, la structure orale dans la prononciation des valeurs numériques et enfin l'acte de leur transcription graphique face au phénomène de la bidirectionnalité alphanumérique. Nous aborderons ces trois

facteurs en présence d'un référentiel culturel multiple pour soulever les anomalies annoncées auparavant concernant les solutions informatiques multilingues.

3.2.1 La calculabilité arithmétique et le mécanisme incrémentiel des chiffres

Sans besoin de revenir sur la structure et le fonctionnement du système numérique décimal à valeurs de positions, et sans évoquer l'argument latin, à notre sens redondant, de positionner les poids les plus importants vers la gauche dans une valeur décimale, il est très clair que le mécanisme incrémentiel des chiffres dans une optique arithmétique décimale (même binaire) respecte encore la directionnalité Droite-Gauche selon le principe d'un accroissement du plus petit vers le plus grand (Unité, dizaine, centaine, millier etc.).

Dans un processus de comptage séquentiel qui constitue la forme de calcul la plus simple, un chiffre évolue à l'image d'un chronomètre digital par une succession de modifications cycliques de l'élément le plus à droite vers l'élément le plus à gauche.

Dans les deux derniers blocs du tableau suivant, le nombre « 1987 » change (évolue) de manière séquentielle par colonne successive de droite à gauche pour arriver à la valeur « 2000 ».



Milliers	Centaines	Dizaines	Unités
0	0	0	1
0	0	0	9
0	0	1	0
0	0	9	9
0	1	0	0
0	9	9	9
1	0	0	0
1	9	8	7
1	9	8	8
1	9	8	9
1	9	9	0
1	9	9	8
1	9	9	9
2	0	0	0

Tableau 52 : Mécanisme incrémentiel des chiffres

Dans des opérations de calculabilité arithmétique plus complexe, le même principe d'évolution droite-gauche est observé. Dans l'exemple suivant, le total d'une somme sous le tableur Excel est obtenu par une progression de calcul et de retenu factoriel de droite vers la gauche. Aucun traitement de gauche à droite n'est observé hormis l'opération de saisie des chiffres standardisée (à tors !) de gauche à droite pour toutes les solutions informatiques dans toutes les langues du monde. L'abstraction et la neutralité linguistique des chiffres en ont fait un modèle unique conçu à l'origine par et pour les concepteurs occidentaux sur une base de confusion directionnelle entre les trois facteurs cités.

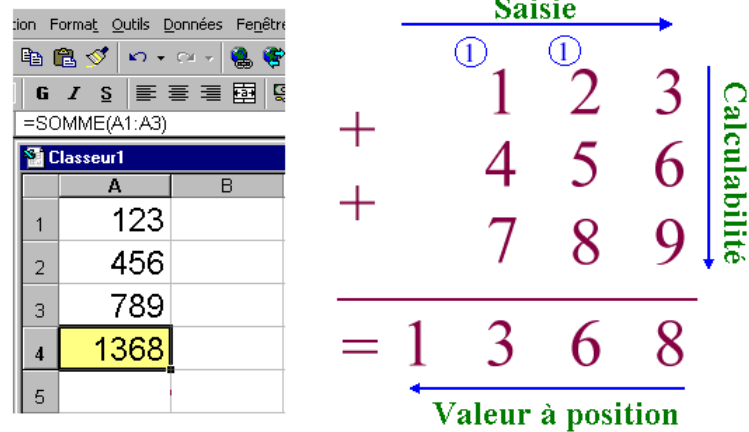


Figure 169 : Calculabilité arithmétique de Droite à Gauche

À notre sens rien n'a forcé, dès le début de la création de ce système numérique, cette directionnalité interne de la calculabilité arithmétique des chiffres sinon l'origine même du système décimal issue d'une culture à écriture de droite à gauche. Nous avons cité dans les paragraphes précédents que les Égyptiens, les Hindous et surtout les Arabes faisaient suivre la composition des chiffres dans la même direction que leurs écritures respectives. Nous évoquerons de nouveau cette harmonie entre chiffres et lettres dans les paragraphes suivants.

À notre sens, la migration du système numérique décimal vers l'occident européen s'est effectuée sans modification du mécanisme de son fonctionnement interne. Seuls l'acte de la transcription graphique et de l'énoncé oral ont été adaptés à la réalité linguistique occidentale. Ceci justifie entre autres le fonctionnement actuel des algorithmes arithmétiques manuels ou automatisés sur un principe de comptage et d'élévation factorielle opposée au sens de l'écriture latine.

3.2.2 L'acte de la transcription graphique des chiffres et l'effet de la bidirectionnalité alphanumérique

En effet, l'acte de la transcription graphique des chiffres isolés ou accompagnés d'un texte dénote d'une conflictive et d'un manque de standardisation notoires essentiellement dans un contexte multilingue.

3.2.2.1 Le contexte latin

Comme nous l'avons déjà mentionné dans l'exemple précédent, le contexte linguistique latin fonctionne dans une opposition directionnelle entre les deux aspects essentiels de la gestion des chiffres à savoir la saisie (transcription graphique) et la calculabilité des valeurs numériques.



Figure 170 : Opposition entre la transcription et les valeurs à positions pour les chiffres en mode latin standard

Cette opposition entre les deux facteurs de saisie et de calculabilité des chiffres en mode latin, crée une autre opposition dans une suite de caractères alphanumériques.

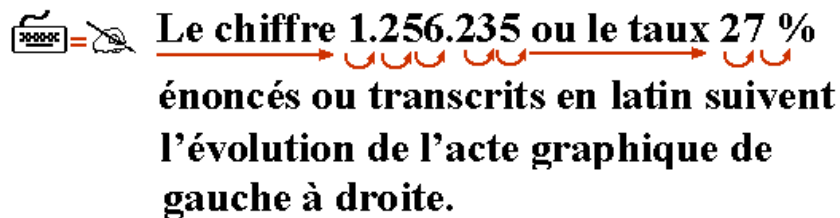


Figure 171 : Accord directionnel entre texte latin et facteur de transcription de chiffres

Seulement, quand le facteur incrémentiel est pris en compte, une opposition d'un autre genre apparaît. Le traitement séquentiel des caractères alphanumériques s'effectue en deux phases opposées : une suite linéaire de gauche à droite pour transcrire les caractères alphabétiques latins et une progression linéaire de droite à gauche pour la disposition des caractères numériques selon le principe des valeurs positionnelles.

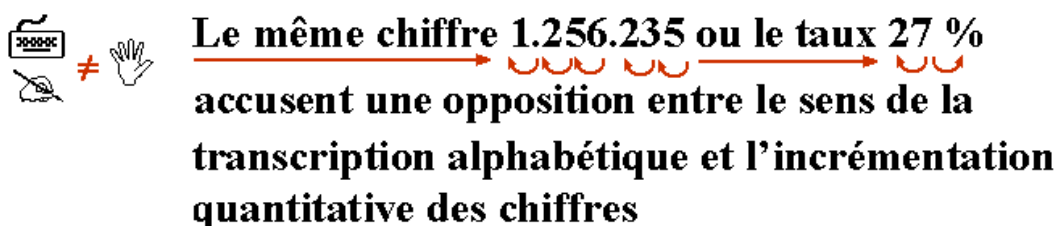


Figure 172 : Opposition directionnelle entre texte latin et calculabilité des chiffres

3.2.2.2 Le contexte arabe

À cette incohérence dans le contexte latin s'ajoute la complexité de la langue arabe vis à vis des langues latines et vis à vis d'elle-même.

Dans la communauté arabophone, la transcription des chiffres n'est pas moins conflictuelle bien qu'elle ait toujours été conforme à l'acte graphique des caractères alphabétiques, c'est à dire de droite à gauche jusqu'à la période coloniale. Il est encore très courant de trouver des anciens

adeptes des écoles coraniques et des cercles littéraires classiques (i.e. Le Azhar en Egypte et la Zitouna en Tunisie) écrire les chiffres et les dates dans une orientation droite-gauche. D'ailleurs la transcription des fractions et des valeurs décimales (virgule flottante) s'effectue dans un texte arabe officiel de Droite à Gauche. La valeur deux tiers (2/3) s'écrirait dans un texte arabe officiel sous forme (3/2). La valeur décimale « vingt-cinq virgule dix » (25,10) s'écrirait en arabe officiel (10,25). Ceci n'est cependant pas commun dans les systèmes d'information d'ordre public (médias, textes juridiques, documentation financière...). La littérature peut être considérée comme le dernier bastion qui garde encore ses origines classiques suite à la nature même des rapports limités qu'elle entretient avec la culture occidentale. Dans ces cercles relativement fermés aux influences technologiques, l'usage des chiffres suit encore les règles de l'arabe standard profond dans le respect de la directionnalité Droite-Gauche en ce qui concerne les trois facteurs de la transcription graphique, la structure orale (voir partie suivante) et la calculabilité arithmétique.

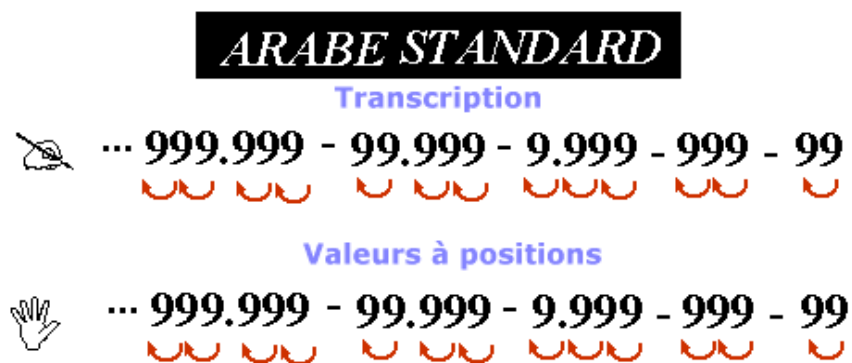


Figure 173 : Concordance entre la transcription et les valeurs à positions pour les chiffres en mode arabe standard

L'effet retour de la migration du système numérique vers l'Occident à engendré la genèse d'un mécanisme hybride de transcription des chiffres. L'acte graphique s'est trouvé ainsi influencé par le mécanisme latin de la transcription de gauche à droite même dans une directionnalité de base de droite à gauche.

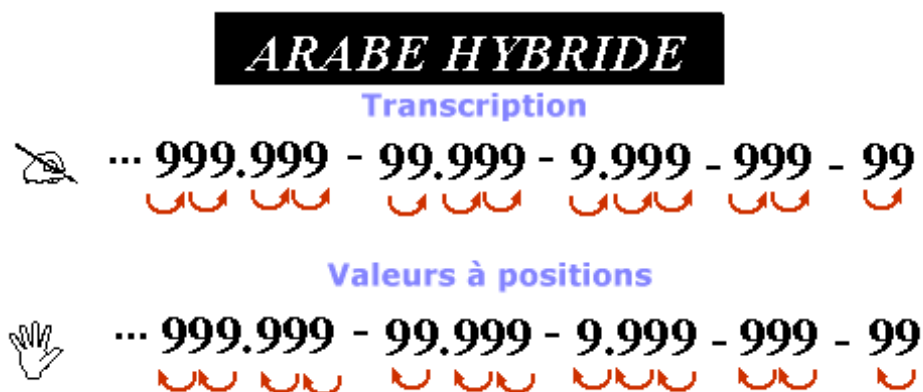


Figure 174 : Directionnalité opposée entre transcription et calculabilité des chiffres en mode arabe hybride

En fait, cet exemple renferme une anomalie latente qui se situe au niveau de la transcription des deux premières colonnes (des unités et des dizaines) d'une valeur numérique quelconque. Les

observations auxquelles nous sommes parvenues concernant ce point particulier, ont démontré que deux attitudes de transcription de ces deux valeurs sont possibles.

Quand la valeur du nombre à transcrire est limitée aux deux colonnes des unités et des dizaines exclusivement, la majorité des personnes composent le nombre dans une directionnalité de droite à gauche. Au contraire, quand le nombre à composer dépasse la valeur des dizaines, l'on tend généralement à composer le nombre dans une directionnalité de gauche à droite. Très peu de personnes composent un nombre de 3 chiffres (ou plus) en commençant à partir de la gauche puis une fois arrivées au niveau des deux dernières colonnes des unités et des dizaines basculent vers la droite pour composer les deux derniers chiffres. De même, un nombre à deux chiffres est généralement transcrit de Droite à Gauche alors qu'un nombre de plus de deux chiffres est plutôt écrit de Gauche à Droite de manière plus fréquente.

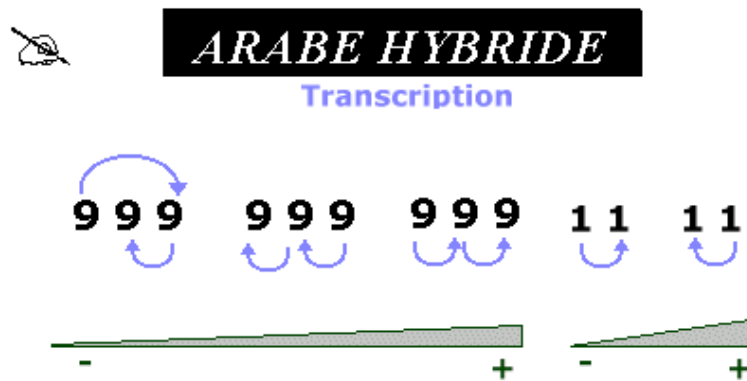


Figure 175 : Polémique des deux colonnes de chiffres aux poids les plus faibles

Cette irrégularité est en réalité associée de façon intrinsèque à la structure orale des valeurs numériques que nous développerons dans les paragraphes suivants.

Cet aspect hybride est plus visible dans une combinaison entre chiffres et texte arabes. Alors que les langues latines ont pu adapter la directionnalité de l'acte graphique des chiffres à celle de leurs écritures alphabétiques, la langue arabe connaît actuellement deux situations combinatoires entre chiffres et lettres ; l'une standard ou littéraire et l'autre hybride ou composée.

يُعتبر الرقم 1.256.432 أو القيمة 27% مطابقاً لفضاً أو نسخاً
لاتجاه الكتابة العربية من اليمين إلى اليسار في إطار لغوي مقنن.

Figure 176 : Le sens de la transcription alphanumérique en mode arabe standard

Dans cet exemple la transcription alphanumérique obéit à une seule directionnalité de Droite à Gauche pour les chiffres et les lettres. C'est la méthode culte des hommes des lettres classiques et des érudits monolingues arabes.

En effet, la culture arabe a toujours respecté sa logique directionnelle droite-gauche même en ce qui concerne la transcription des chiffres. Dans la littérature classique, les chiffres étaient écrits

et prononcés dans un esprit arithmétique qui se basait sur la logique de l'évolution unitaire, décimale... La valeur « 1998 » est à titre d'exemple écrite (à la main bien entendu) dans une suite successive du geste graphique (Droite-Gauche) qui passe du dernier caractère du dernier mot écrit (le cas où le chiffre fait partie d'une chaîne de caractères) vers l'emplacement situé immédiatement à sa gauche afin de continuer avec le chiffre « 8 » suivi des chiffres « 9 », « 9 » et « 1 ». La prononciation du même chiffre est également accomplie dans cette logique d'évolution des unités vers les dizaines, les centaines et les milliers.

لكن لا يُعتبر الرقم 1.256.432 أو القيمة 27% مطابقان لفظاً أو نسخاً
لا تَجَاهِ الكِتَابَةَ العَرَبِيَّةَ مِنَ الْيَمِينِ إِلَى الْيَسَارِ فِي إِطَارِ لُغَوِيِّ هَجِينِ.

Figure 177 Sens de la transcription alphanumérique en mode arabe hybride

Par contre, dans cet autre exemple, l'influence occidentale se fait sentir au niveau de la transcription des chiffres qui s'aligne au modèle latin de gauche à droite. C'est la méthode de l'arabophone commun, des mass-media et des secteurs scientifiques et technologiques très en relation avec la culture occidentale.

En effet, les répercussions de l'effet retour du système numérique de l'Occident vers l'Orient durant la période coloniale ont fait que le système numérique avait été réimplanté dans son contexte d'origine avec les modifications qui lui ont été apportées dans le contexte linguistique latin. Il est ainsi devenu plus courant aujourd'hui d'écrire les chiffres dans un contexte arabe selon le modèle occidental. Il est même pire d'observer qu'aucun modèle d'écriture des chiffres dans un contexte arabe n'est encore défini comme standard de fait. En terme d'interface Homme-Machine, la solution informatique standard basée sur le modèle latin s'établit désormais comme aiguilleur vers un modèle unique, le modèle à directionnalité Gauche-Droite. Devant la machine toutes les langues sont égales en ce qui concerne la transcription des chiffres.

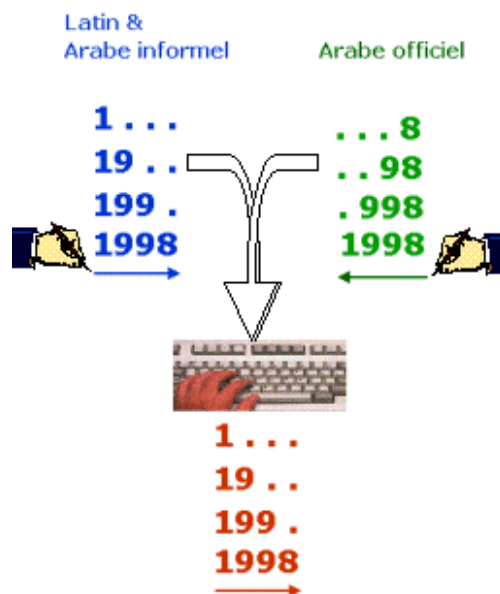


Figure 178 : Modes de transcription graphique des chiffres

Les dates en langue arabe attestent de la même polémique d'usage. La date « 11 février 1998 » est une suite de valeurs qui à vue d'œil respectent la logique évolutive du plus petit au plus grand; en l'occurrence le jour, le mois et l'année. Seulement, si nous observons avec un soubassement algébrique la transcription des valeurs de cette date nous remarquerons qu'au niveau de la valeur jour, les dizaines se placent dans cette orientation latine de gauche à droite, avant les unités ; chose qui contredit la dite logique évolutive du plus petit au plus grand. Par contre si nous observons, selon le même intérêt arithmétique, la date arabe « 11 فيفري 1998 », nous constaterons qu'elle respecte cette même logique évolutive sur un plan sémantique et algébrique à la fois. Partant de droite à gauche, nous avons dans une suite successive l'unité puis les dizaines pour les jours, ensuite le mois et enfin l'année elle même commençant par les unités et poursuivant par les dizaines, les centaines et les milliers.

3.2.3 La structure orale dans la prononciation des chiffres :

Le facteur de l'oralité joue un rôle important dans cette polémique des directions opposées. Il y constitue d'ailleurs un facteur déterminant pour le mécanisme de la transcription graphique des chiffres.

Comme nous l'avons déjà mentionné avec l'exemple des chiffres en mode latin, l'énonciation d'un chiffre correspond à sa directionnalité graphique mais s'oppose à sa structure arithmétique interne. Le chiffre « 1 256,235 » est prononcé « Un Million (1) Deux cent (2) Cinquante (5) Six mille (6) Deux cent (2) Trente (3) Cinq (5) » respectant ainsi une directionnalité gauche droite intégrale.

Le cas de la langue arabe présente par contre une polémique similaire au mode de transcription graphique des chiffres. Une opposition entre le discours littéraire et la langue populaire persiste et s'accentue davantage sous l'influence d'un courant culturel et technologique occidental de plus en plus fort.

En arabe littéraire, toute valeur numérique est prononcée conformément à sa structure arithmétique interne ainsi qu'au mode de sa transcription graphique. Par conséquent, un érudit « Zitounien » ou un « Fakih Azhérien » prononcerait la date du 28 Mai 1998 comme suit : « Huit et Vingt Mai, Huit et Quatre vingt Dix et Neuf Cent et Mille ». Cette pratique constitue jusqu'à nos jours l'une des marques de distinction de la bonne maîtrise de la langue arabe.



Figure 179 : Prononciation littéraire des chiffres

Seulement, suite aux interférences culturelles elles-mêmes dues en grande partie au phénomène retour que nous avons cité auparavant, les chiffres et les dates s'alignent depuis l'époque coloniale au modèle latin.

Dans notre analyse de la structure orale des chiffres, nous avons également détecté une anomalie identique à celle relative au mode de transcription graphique des deux valeurs des unités et des dizaines soulevées dans le point précédent.

Reprenant l'exemple précédent, toute personne, autre que les adeptes de la littérature classique, le prononcerait comme suit : « Huit et Vingt mai, Mille et Neuf cents et Huit et quatre-vingt-dix ».



Figure 180 : Prononciation hybride des chiffres

La seule distinction dans cette polémique entre le facteur de la transcription graphique et l'énonciation (oralité) des chiffres en mode arabe hybride est le consensus implicite et général de toute la communauté arabophone pour prononcer les unités et les dizaines dans une directionnalité de droite à gauche alors que l'acte graphique reste tributaire de beaucoup de facteurs personnels de styles, d'apprentissage et d'habitude.

En réalité ce phénomène ne se limite pas aux valeurs des unités et des dizaines mais se situe plutôt au niveau des deux premières positions (à partir de la droite) dans un groupe de trois. Ce n'est qu'à partir de la troisième position, en l'occurrence la valeur « 100 » et ses multiples, que l'inversion dans l'ordre de prononciation prend forme. Ainsi la valeur « 21 » est prononcée « Un et Vingt » ; la valeur « 121 » est prononcée « Cent et Un et Vingt » ; la valeur « 1 121 » est prononcée « Mille et Cent et Un et Vingt » ; la valeur « 21 121 » est prononcée « Un et Vingt Mille et Cent et Un et Vingt » ; la valeur « 121 121 » est partagée en deux blocs pour être prononcée « Cent et Un et Vingt Mille et Cent et Un et Vingt »

121 121	21 121	1 121	121	21

Figure 181 : Structure orale des nombres en mode arabe hybride

Ce paradoxe constituerait à notre sens un thème de recherche innovateur dans le domaine de la sociolinguistique.

En poussant le questionnement de ce phénomène particulier aux valeurs unitaires et décimales à son paroxysme sociolinguistique, nous avons découvert une ressemblance non moins étonnante entre le mécanisme de l'articulation des chiffres en mode arabe hybride et le système de prononciation des chiffres en langue ... allemande (une piste de recherche supplémentaire à déblayer !?).

En allemand, et contrairement à toutes les langues latines et germaniques (un degré de ressemblance est quand même observé avec la langue anglaise !), les unités et les dizaines sont prononcées dans une directionnalité de droite à gauche. Une fois passée aux valeurs supérieures,

la directionnalité est partiellement inversée vers la gauche-droite, exactement comme le modèle arabe hybride.

Le tableau suivant établit une analogie entre les structurations verbales de certains repères numériques choisis dans quatre langues à savoir l'allemand, l'arabe hybride, le français et l'anglais.

Valeur Numérique	Allemand	Arabe ^(*)	Français	Anglais
1	Eins	Un	Un	One
10	Zehn	Dix	Dix	Ten
11	Elf	Un et Dix	Onze	Eleven
12	<u>Zwölf</u>	<u>Deux</u> et Dix	<u>Douze</u>	<u>Twelve</u>
19	<u>Neun</u> zehn	<u>Neuf</u> et Dix	Dix <u>Neuf</u>	<u>Nineteen</u>
20	<u>Zwanzig</u>	Vingt	Vingt	<u>Twenty</u>
21	Eins Und Zwanzig	Un et Vingt	Vingt et Un	Twenty One
99	Neun Und Zwanzig	Neuf et Quatre Vingt Dix	Quatre Vingt Dix Neuf	Ninety nine
100	Ein Hundert	Cent	Cent	One hundred
101	<u>Ein Hundert Eins</u>	<u>Cent et Un</u>	<u>Cent et Un</u>	<u>One Hundred One</u>
121	Ein Hundert Eins Und Zwanzig	Cent et Un et Vingt	Cent <u>Vingt et Un</u>	One hundred <u>Twenty One</u>
1998	Ein Tausend neun Hundert <u>Acht Und Neunzig</u>	Mille et Neuf Cent et <u>Huit et Quatre Vingt Dix</u>	Mille Neuf cent Quatre Vingt Dix Neuf	One Thousand Nine Hundred Ninety Eight

(*) = Arabe hybride translittéré pour des raisons de communicabilité et de comparaison avec les équivalents latins.

Tableau 53 : Analogie entre les structures morphologiques des chiffres en allemand, arabe hybride, français et anglais

Parmi les constations premières qui surgissent à travers ce tableau, notons surtout :

- ❑ Les 4 langues considérées se rallient en terme de structure morphologique des chiffres par paires de langues opposées, la première constituée par les langues arabe et allemande et la deuxième composée par les langues française et anglaise ;
- ❑ La langue anglaise constitue pourtant une exception dans le sens où la structure verbale des chiffres y est inspirée des deux modèles français et allemand à la fois. Les chiffres entre 0 et 19 sont prononcés sur le modèle allemand et arabe à savoir une combinaison « Unité-Dizaine ». Par contre les chiffres entre « 21 » et « 99 » sont d'une inspiration française à savoir « Dizaines-Unités » ;
- ❑ Les 4 langues s'accordent pour l'usage d'une structure identique des chiffres entre 0 et 19 en tant qu'une combinaison « Unités et Dizaines ». Toutefois, le français inverse cette structure à partir de la valeur « 17 » pour devenir une combinaison « Dizaines-Unités ». La valeur « 16 » est une combinaison d'un « Sei », un pseudo « Six » et d'un « ze », un pseudo « Dix ». La valeur « 17 » est par contre une combinaison claire entre un « Dix » et un « Sept ».
- ❑ L'adjonction d'une valeur d'un poids supérieur (centaines, milliers etc.) à cette rangée de valeurs n'affecte pas cette structure combinatoire ;

S'il est vrai que ce phénomène de directionnalité des chiffres n'est pas très soulevé dans les milieux de travail multilingue, son analyse épistémologique fait resurgir cette série d'antagonismes et de similitudes qui dénotent une fois de plus de la confluence culturelle universelle, des influences et des interférences linguistiques multidirectionnelles. La question

posée serait toujours celle du rapport entre l'efficacité d'un système opérationnel et sa dimension culturelle. Il s'agit là de deux approches analytiques différentes qui s'opposent mais qui ne s'annulent pas. Notre choix s'est effectué dans cette étude sur la deuxième approche analytique comme facteur tributaire du contexte de l'information et de la communication plutôt que de l'analyse des systèmes.

CONCLUSION GÉNÉRALE

Depuis la problématique annoncée au début de ce travail, nous avons voulu dissiper une confusion potentielle relative à la portée de notre intitulé de recherche qui cerne à priori le domaine des interfaces Homme-Machine. Nous avons canalisé l'évolution de l'analyse sur un terrain relativement parallèle par rapport au noyau dur des interfaces qui couvre toujours l'un des deux concepts d'évaluation ou de conception des IHMs. L'univers des caractères et de leurs représentations codées nous a convoité par l'originalité de la problématique que nous nous sommes posée : un multilinguisme lourd qui puise ses origines dans deux facteurs déterminants, à savoir la bidirectionnalité et la diversité dans les formes des glyphes. Ces deux facteurs ont toujours fait des langues à écritures non latines, des langues conflictives avec le standard latin et en particulier le modèle anglo-saxon.

Notre intrigue n'émanait pas à juste titre de cette opposition et cette divergence dans la directionnalité de l'acte graphique ou de la forme des glyphes d'une langue ou d'une autre. Le plus frappant pour nous c'était de voir que les solutions et les modèles adoptés pour résoudre l'isolement technologique des langues dites vernaculaires partaient d'une assise latine et se faisaient « enrober » dans une parure linguistique locale. Notre sensibilisation à cet état de fait émane d'un vécu dans lequel le référent socioculturel se heurte au techno-scientifique sans pour autant interrompre la symbiose qui aboutit à un standard de fait : les interfaces multilingues ad hoc.

Pourtant la question qui reste toujours à poser, même à terme de cette exposition des aspects du multilinguisme lourd arabe/latin, est de savoir le rapport direct que ces facteurs que nous avons identifiés de "catalyseurs" ont avec les Interfaces Hommes Machine.

La réponse est à doubles penchants. Après avoir souligné l'importance des traitements sur la langue en ce qui concerne le codage et la représentation interne dans le processus de dialogue Homme-Machine, deux situations peuvent être analysées comme continuité ou comme complément à une étude ergonomique des interfaces home-machine : une approche orientée utilisateur et une approche orientée système.

La première approche consiste à aborder la dimension psychocognitive de l'utilisateur en tant que vis à vis dynamique qui dialogue avec le système d'information à travers une interface machine. *« Pour garantir l'ergonomie des interprétations construites par la machine, c'est-à-dire, leur conformité aux attentes des utilisateurs, le fonctionnement du système mis en œuvre doit être analogue à celui de la cognition humaine. Ainsi, que l'on veuille communiquer avec la machine ou communiquer avec d'autres individus ou d'autres systèmes par son intermédiaire, elle a besoin de facultés lui permettant de dialoguer efficacement »*^[1].

Or, selon l'approche orientée système, les environnements informatiques changent et évoluent à une vitesse telle que les modalités de conception d'interfaces Homme-Machine, qui normalement devraient être éprouvées et maîtrisées par le biais de spécialistes en IHM (ergonomes entre autres), sont progressivement tributaires de « *la pression compétitive entre constructeurs de logiciels* »^[2]. Fanchini signale à cet égard que certains stéréotypes d'interfaces s'établissent comme modèle de référence clé en mains standard et univoque. C'est le cas des systèmes à fenêtres et les interfaces graphiques du type Windows. Seulement, si cette approche a

[1] Gérard SABAH. Groupe Langage et Cognition. <http://www.limsi.fr/Individu/gs/GroupeLC/Theme3.html>

[2] H. FANCHINI. Bilan de cinq années de pratique en conception des IHM. Analyse d'un panel de projets informatiques. *Actes du XXIX^e Congrès de la Société d'Ergonomie de Langue Française (SELF)*, Paris. 1994. <http://www.fdn.fr/~hfanchini/PUB/94BIL/94BIL.HTML>

facilité l'apprentissage des nouveaux logiciels conçus sur les mêmes principes appuyés à leur tour par la dimension « auto-portante » sur le plan didactique, « *la convivialité de bon aloi est parfois trompeuse* ».

En effet, les interfaces et les applications se complexifient de plus en plus et se truffent de fonctionnalités multiples de sorte que les usagers ne parviennent plus à utiliser toutes les fonctionnalités qui sont mises à leur disposition. La convivialité aurait ainsi permis à l'utilisateur de maîtriser certains outils et certaines fonctionnalités sans pour autant avoir une connaissance nécessaire du bien fondé de ses interventions ni de la « traçabilité » des opérations réalisées par le système. Ceci est d'ordre à se répercuter inéluctablement sur la véritable dimension ergonomique de l'interface.

Dans le cas concret de notre étude relative à l'étude des interfaces multilingues de communication Homme-Machine, une première observation met en évidence le maintien des standards de morphologie et d'architecture des solutions monolingues latines avec adaptation de l'habillage par adjonction de zones spéciales gérant la multiplicité linguistique.

Une question de toute première importance est de savoir si l'adjonction de ces zones et boutons spéciaux pour la gestion de l'aspect multilingue des interfaces Homme-Machine est accomplie sur une base expérimentale ou esthétique. Car si le maintien des aspects de la même interface latine pour concevoir l'alternative multilingue est justifié par l'harmonie des plates-formes de travail et la fidélisation de l'utilisateur, les solutions apportées pour le choix et le changement de langues d'interfaces varient en aspects (menu de commandes, raccourcis claviers, boutons de raccourci, zone hypertexte, liste à cocher, bouton radio ...) et en fonctionnalités (gestion de la bidirectionnalité de l'écriture et du sens du texte ; la désignation et l'invocation des pages de codes ; les effets de justification de texte ...).

Il est évidemment hors de question pour le moment d'aller contre le « *confort hérité* » des interfaces graphiques courantes tellement elles sont bien enracinées dans le mental des usagers, matricées et mises en forme dans leur pensées et proposer une autre catégorie d'interfaces dans un nouveau « look » et une architecture différente. L'époque des applications à commandes procédurales et les systèmes à menus déroulants est bien révolue (seuls quelques vestiges de l'histoire végètent encore grâce à la chirurgie esthétique du revamping !). Il faudrait attendre une nouvelle génération d'interfaces qui se libéreraient des contraintes de l'interaction cliquable (souris, claviers) et proposer des alternatives vocales ou même olfactives (Sic!). L'essentiel est de définir des interfaces « simples et non simplistes » capables de respecter la communicabilité réciproque : provoquer chez l'utilisateur un mécanisme de « *construction d'une représentation opératoire adéquate, pas forcément exacte mais valide* » et permettre du côté de l'interface de refléter le mécanisme du fonctionnement interne de la machine pour que l'utilisateur puisse être averti des détails sur les tenants, les aboutissants et les limites de ses interactions avec elle. En d'autres termes, il s'agit de sonder à quelle profondeur les solutions existantes d'interface Homme-Machine combinent-elles le soubassement cognitif et les référents socioculturels de l'utilisateur avec la morphologie et les architectures des interfaces sur lesquelles il intervient. L'objectif est de mesurer la concordance entre les représentations potentiellement existantes dans l'imaginaire de l'Homme et les représentations symboliques au niveau de l'interface. Car en fait, « *face aux changements technologiques incessants, le seul objet de recherche qui demeure stable dans le temps concerne le fonctionnement de l'esprit humain* »^[3].

[3] Henri FANCHINI. Bilan de cinq années de pratique en conception des IHM. Analyse d'un panel de projets informatiques. Op.Cit.

Jusqu'à ces dernières années, la majeure partie des projets informatiques visaient à informatiser et automatiser des tâches manuelles existantes. L'on se situait alors dans une *problématique de transposition*, avec pour paradigme la recherche des invariants.

Il s'avère désormais que certains projets s'attachent dans un même mouvement à informatiser des tâches en même temps qu'ils les créent. L'ergonome est alors confronté à une *problématique de recomposition*, voire de *création*, dans lesquelles les situations de référence ne sont pas données, au mieux doivent être construites à partir de fragments d'activité, d'organisation, d'utilisateurs et d'outils, et au pire sont inaccessibles (les tâches, les utilisateurs, et même la structure organisationnelle n'existent pas).

Dans les méandres de ces turbulences conceptuelles, architecturales et ergonomiques des interfaces, nous avons essayé à travers notre présente étude d'aller à contre courant des faits établis afin de démontrer certains paradigmes et aboutir aux conclusions génériques suivantes :

- ❑ Tous les systèmes d'information et de communication unilingues ou multilingues que nous avons pris comme échantillons fonctionnent à base d'une série d'algorithmes divergents et irréguliers sur la base du traitement interne des données (le codage, le tri et le traitement Bidi). Nous avons démontré à travers des exemples pris dans des situations et des contextes réels, l'opposition flagrante entre les traitements logiques internes des systèmes et l'affichage ou le rendu du texte au niveau des IHMs ;
- ❑ La contextualisation des systèmes de tri selon les plates-formes MS-Dos ou Windows des applications latines ou multilingues arabe/latin, est d'ordre à alourdir les systèmes de traitement par l'adjonction de programmes, d'algorithmes et de locales supplémentaires de filtrage, de conversion et de reformatage des données ;
- ❑ La bidirectionnalité est un facteur encore non résolu au niveau des interfaces Homme-Machine multilingues. L'unicité directionnelle des chiffres de Gauche à Droite sans considération de l'analyse contextuelle est une source énorme de confusion au niveau de l'interface pour des utilisateurs non avertis. Cet état de fait est compliqué davantage dans des situations de traitement des bris et des incises. Notre analyse nous a cependant permis de constater une analogie, restée encore énigmatique pour nous, entre les deux langues arabe et allemande au niveau de la structure arithmétique et du traitement des chiffres ;
- ❑ L'évolution dans l'intégration de la langue arabe dans le contexte informatique et plus particulièrement dans les systèmes d'information et de communication multilingues est certes en progression constante. Seulement, nous avons pu observer que cette évolution ne couvre généralement que le champ d'action de l'utilisateur. Elle n'a pas atteint un degré majeur de maturité comme langue qui touche les zones systèmes. Si certains parmi ces derniers qui fonctionnent en mode monoposte ont pu résoudre cet aspect de façon locale par l'arabisation des systèmes d'exploitation des machines (MS-DOS, Windows), l'hégémonie demeure par contre latine et plus particulièrement anglo-saxonne dans la manipulation des données systèmes (URIs, paramétrage des protocoles ...) dans les univers des réseaux larges distances.
- ❑ La diversité des systèmes de codage constituera encore et pour assez longtemps un blocage considérable vers l'i18n. L'avènement d'Unicode et de ISO/IEC 10646 restera à notre sens une alternative incontournable bien qu'encore lointaine vers la transparence linguistique universelle. Les techniques d'extension de codes selon les normes ISO/IEC 2022 et assimilées resteront encore en vigueur pour un temps assez considérable ;
- ❑ L'usage de la langue arabe dans les applications basées sur les nouvelles normes d'Unicode et d'ISO/IEC 10646 n'est pas encore adopté unanimement à l'instar des autres langues non-latines comme le cyrillique ou les langues asiatiques. Les applications

multilingues les plus utilisées dans le contexte monoposte ou en réseau résolvent le problème de la langue arabe par la rédaction des versions appropriées et propriétaires clés en main pour la langue arabe au lieu d'en faire une alternative linguistique gérée par Plug-Ins. Les cas de Netscape et de Microsoft en sont des exemples concrets ;

Notre contribution pour faire face à ces zones de difficulté et d'anarchisme dans les pratiques et les usages de la langue arabe a versé d'une part dans le soulèvement des problématiques peu documentées mentionnées ci-dessus et la proposition des quelques alternatives théoriques selon une logique d'équité linguistique, de structure fonctionnelle des systèmes d'information multilingues conformes à des pré-requis culturels, techniques et linguistiques de la langue arabe.

Nos propositions étaient d'une part structurelles concernant le traitement du caractère arabe au niveau des tables des codes et d'autre part fonctionnelles concernant son usage comme langue système dans l'identification des ressources et le paramétrage des protocoles dans l'environnement informationnel et communicationnel des réseaux mondiaux. Nous avons ainsi proposé les recommandations suivantes :

- ❑ La codification des glyphes et des caractères doit être homogène au niveau des caractères simples et composés. Le problème du caractère spécifique « Lam-Alef », des paires des caractères à pivotement contextuels est une source majeure de divergence entre systèmes de codage. Réécrire les tables des codes des caractères arabes selon une répartition uniforme est devenu nécessaire après l'évolution enregistrée dans les techniques de codage et de traitement des caractères codés sur plusieurs octets. Le « prétexte » de l'exiguïté des espaces réservés dans les tables des codes à 7 ou 8 bits, que les anciens concepteurs des tables des codes avançaient toujours comme argument, est actuellement résolu par le grand nombre d'espaces non encore attribués dans les matrices Unicode et ISO/IEC 10646.
- ❑ Les caractères arabes doivent être présents dans les zones systèmes des machines et protocoles d'information et de communication. Nous avons abordé dans ce sens le phénomène du multilinguisme des URIs et de l'identification des DNS sous Internet et nous avons proposé un modèle de codification qui se base sur le codage et le décodage des caractères multilingues selon un mécanisme numérique. Partant du principe de l'abstraction linguistique des chiffres, chaque caractère de n'importe quelle langue sera converti en une valeur numérique unifiée transmissible sur les réseaux et reconvertie à son état initial par les protocoles et les algorithmes de conversion appropriés. Nous admettons que cette proposition nécessite un volet expérimental intense pour la juxtaposer avec la solution parallèle en cours d'implantation pour l'usage d'UTF-8 en format %HH ;
- ❑ En l'absence de solutions Unicode généralisées, nous avons proposé la manière comment consolider l'intégration de la langue arabe dans les processus de la négociation linguistique dans les contextes des systèmes ouverts et distribués par l'interposition des serveurs de glyphes, le recours aux systèmes de l'hypertexte et de la traduction parallèles. Notre intervention sur ce point n'est pas innovatrice. Elle consiste uniquement à expliquer les modalités de fonctionnement de ces systèmes à base des caractères arabes. L'absence de supports logistiques nécessaires (programmes de traduction automatisée en mode parallèle, plates-formes expérimentales distribuées ...) nous ont contraint à aborder la thèse d'un point de vue conceptuel et schématique ;
- ❑ Le phénomène majeur que nous pensons avoir contribué à exposer ses différentes polémiques d'usage et sans doute le phénomène des chiffres dans un système positionnel. Le monde des chiffres dans un système décimal laisse entrevoir une

origine directionnelle Droite-Gauche mais une confusion énorme au niveau de leur transcription, de leur prononciation et de leur calculabilité qui varient selon les contextes. Malgré cette directionnalité originale encore appliquée par tous les algorithmes de calcul, tous les systèmes informatiques actuels appliquent la directionnalité opposée (Gauche-Droite) dans l'acte de la transcription graphique. L'antagonisme systémique des chiffres se révèle ainsi enraciné dans les deux systèmes unilingues et multilingues. Rétablir le traitement des chiffres dans un contexte arabe à leur directionnalité d'origine (Droite-Gauche) représenterait à notre sens une tâche relativement simple sauf complication de l'envergure des changements à effectuer sur les systèmes ad hoc construits sur le principe de l'opposition entre le traitement interne et le rendu visuel. La simplicité (relative !) émane des deux facteurs essentiels à savoir l'agissement limité sur la modification des codes internes de la directionnalité gauche-droite héritée des systèmes latins et la conformité que ce changement assurerait avec le mécanisme de fonctionnement des algorithmes internes de la calculabilité arithmétique ;

En définitive, nous admettons qu'il est encore très prématuré de parler d'une tendance en cours de concrétisation vers la révocation des systèmes en place conformément à notre approche de la question du multilinguisme arabe/latin car, comme nous l'avons cité un peu plus haut, il est encore très tôt pour aller à l'encontre d'un « confort hérité » des interfaces conçues dans une logique de marché monopolisante où la seule logique économique façonne et conditionne les goûts et les habitudes aux dépens des soubassements et des origines culturelles et linguistiques millénaires. Est-ce là une des formes de la mondialisation sur le modèle occidental de cette fin de siècle ? Le concept de la localisation, sœur jumelle de ce phénomène de globalisation, est pourtant omniprésent dans tous les discours politiques, économiques et culturels actuels : s'agit-il d'une garantie pour un univers *riche par sa diversité culturelle* ou est-ce un leurre qui camoufle l'effet de l'économie d'échelle et la culture unique. A bon lecteur, bon suivi.

.../...

BIBLIOGRAPHIE UTILISÉE

1. ABBA L. etc. Principe pour l'établissement d'une interconnexion entre les réseaux informatiques des pays africains. Actes de la réunion de lancement du réseau d'échanges d'informations entre institutions en Afrique. Dakar, Sénégal. 27-29 février 1992.
2. ABOULHAYJA Ahmed Muwasafat wa makais litaarib al maaloumatia=Normes et standards pour arabiser l'informatique. In : Istikhdam Allougha Al Arabia fi al Maaloumatia = Usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.150-188.
3. About the HTML 4.0 specification. W3C Working Draft 17-Sep-1997. <http://www.w3.org/TR/WD-html40-970917/>
4. ADAKOLE Ochai. The emerging information society. International Library Review, N° 16, 1984 ; pp.367-372
5. ANDRE Jacques et GOOSSENS Michel Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. In : Cahiers GUTenberg n°20-mai 1995.pp.1-53.
6. ASMO. Jeux de caractères arabes/latins codés à 8 éléments pour l'échange de l'information : Normes arabe n° 708 - 1988. Ligue des états arabes. 6p.
7. ASMO. Traitement de l'information - Jeux de caractères arabes codés à 8 éléments pour l'échange de l'information : Norme arabe n° 662 - 1985. Ligue des états arabes. 9p.
8. BARTHES Roland. Théorie du texte. In : Encyclopædia Universalis. France S.A.1995. 22-372 a
9. BASTE Sylvie. Recherche sur la normalisation internationale du codage des caractères codés sur plusieurs octets. Novembre 1997.
10. BEN AHMED Mohamed. Création culturelle et nouvelles technologies de l'information. Al Madar (Cité des sciences. Tunisie). N°3, 1995 ; pp.21-32
11. BENVENISTE Émile. Problèmes de linguistique générale. 2 vol., Paris : Gallimard, 1966-1974.
12. BERNERS-LEE, T., FUIELDING R., FRYSTYK H. Hypertext Transfer Protocol -- HTTP/1.0., RFC 1945 MIT/LCS, UC Irvine, May.1996 <http://www.ics.uci.edu/pub/ietf/http/rfc1945.html>
13. Bibliorom Larousse. Version 1.0. Microsoft Corporation & Liris Interactive. 1996
14. BOURBONNAIS Jean, YERGEAU François Languages on the Internet. Alis Technologies inc. Montréal. http://www.crim.ca/inet96/papers/a5/a5_3.htm
15. BRAND Guy. Qu'est-ce que MIME ? 17 janvier 1995 (Révisé le 4 juin 1997). <http://wwwchimie.u-strasbg.fr/membres/GB/MIME.html>
16. BRETON Philippe. Une histoire de l'informatique. La Découverte, 1990.
17. BRYAN Martin Web Internationalization & Multilingualism Symposium. Séville : WC3, 20-22 Novembre 1996. <http://www.sil.org/sgml/seville-w3c-int.html>

18. CACALY Serge Le paysage documentaire de demain ou le triomphe des interfaces. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992 ; pp.177-182.
19. CAIN Jack. Linguistic diversity, computers and Unicode. <http://www.idrc.ca/library/document/netpac/abs21.html>
20. CANDELIER Michel, ESTIVAL Robert, GARNIER Yves Pour une archéologie schématique de la communication. In : Schéma et schématisation. N°19, 1983. ISSN : 0586-7606. Pp.9-20.
21. CARRASCO BENITEZ Manuel Tomas. On the multilingual normalization of the Web. Third International World-Wide Web Conference. 10-14 Avril 1995. Darmstadt, Germany. <http://www.sil.org/sgml/carrasco.html>
22. CARRASCO BENITEZ Manuel Tomas. WInter (Web Internationalization & Multilingualism). WInter Group, 1996. <http://www.crpht.lu/~carrasco/winter/draft0.html>
23. CASSEN Bernard. Le Tout-anglais n'est pas une fatalité. Le Monde Diplomatique : Manière de voir, octobre 1996, pp.81-82
24. Character Sets - ISO/IEC 2022. EWOS Guide to Open Systems Specifications. <http://www.ewos.be/tg-cs/gis2022.htm>
25. CHRISTIN Anne-Marie, BALPE Jean Pierre. L'écriture. In : Encyclopædia Universalis France S.A. 1995. 7-914 a
26. Comité Consultatif International Télégraphique & Téléphonique <http://www.sandybay.com/pc-web/CCITT.htm>
27. Common Gateway Interface : Mécanismes qui permettent aux programmeurs du Web d'interagir avec des programmes s'exécutant sur un serveur HTTP.
28. COUTAZ J., BALBOS S. Évaluation des interfaces utilisateur : taxonomie et recommandations. Actes des sixièmes journées sur l'ingénierie des interfaces Homme-Machine, IHM'94, 211-218, Lille, 8-9 décembre 1994.
29. CURTIN B. Internationalization of the File Transfer Protocol. FTPEXT Working Group. Ineternet Draft, juin 1998. <http://..draft-ietf-ftpext-intl-ftp-05.txt>
30. DAHMANE Majid. Information scientifique et technique (IST). Communication et langages. N°90, 1991 ; pp.104-105.
31. DALBIN Sylvie. Interfaces dans les systèmes d'aujourd'hui. In : Christian Bornes. Dir. Interfaces Intelligentes dans l'information scientifique et technique. INRIA, 1992. pp.133-173.
32. D'ALEMBERT Jean Le Rond. Discours préliminaire. In : Encyclopédie. Vol.1, 1751.
33. DANIEL Andrew. How to make a Multilingual Web server. <http://vancouver-webpages.com/multilingual/howto.html>
34. DE NOBLET J. La technique dans la culture. Esprit. Octobre 1982.
35. DE ROSNEY Joël. Ce que va changer la révolution informationnelle. Le Monde Diplomatique, août 1996 ; p.19

36. DE SAUSSURE Ferdinand. Cours de linguistique générale. Edition critique préparée par Tullio De Mauro. Payot. Paris : 1986.
37. DE SMEDT Thierry. Introduction aux technologies de l'audiovisuel : textes et schémas aux passages les plus complexes du cours de technologies des communications. Bruxelles : UCL Département de communication, 1996.
38. EVEZE Jean. L'information scientifique et technique : produit scientifique ou discours sur la science ? Communication et langages. N°90, 1991 ; pp.95-103.
39. DILLON Martin. Interfaces for information retrieval and online systems : the state of the art. New York, London : Greenwood Press, 1991. 337 p.
40. DORDICK Herbert. Information society indicators : description, measurement, prediction. In : Information societies : comparing the Japanese and American experiences. Seattle : International Communication Centre, 1978.
41. DOSA Marta. Information Transfer as Technical Assistance for Development. Journal of the American Society for Information Science. Vol.36, n°3, 1985 ; pp.146-152.
42. DUCASSE Roland. Modèle Client-Serveur et systèmes d'information distribués : éléments pour une reingénierie des services d'information scientifique et technique / Roland DUCASSE. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. p.136.
43. DUERST Martin. Archives du forum « URL Internationalization ». Lundi 24 février 1997. <http://www.acl.lanl.gov/URI/archive/uri-96.messages/0010.html>
44. DUERST Martin. Internationalization of Domain Names. University of Zurich. Décembre 1996. <http://www.ifi.unizh.ch/groups/mml/mduerst/urli18n.html>
45. ERICKSON Janet C. Options for Presentation of Multilingual Text : Use of the Unicode Standard. <http://dns.hti.umich.edu/htistaff/pubs/1997/janete.01/>
46. FANCHINI H. Bilan de cinq années de pratique en conception des IHM. Analyse d'un panel de projets informatiques. Actes du XXIX^e Congrès de la Société d'Ergonomie de Langue Française (SELF), Paris. 1994.
47. FEVRIER James G. Histoire de l'écriture. Paris : Payot, 1959. p. 273.
48. FLAJOLET Philippe. Algorithmique. In : Encyclopædia Universalis France S.A. 1995. 1-814 b
49. FLICHY Patrice. La question de la technique dans les recherches sur la communication. Réseaux Communication Technologie Société. N°50, 1991 ; pp.51-62.
50. GANS Paul J. The Medieval Technology Pages : Arabic Numerals. <http://scholar.chem.nyu.edu/technology.htm>
51. GETTYS Irvine J., MOGUL J., FRYSTYK H., BERBERS-LEE T. Hypertext Transfer Protocol -- HTTP/1.1. Request for Comments: 2068. Network Working Group. Janvier 1997. <http://www.ics.uci.edu/pub/ietf/http/rfc2068.txt> ou <http://ds.internic.net/rfc/rfc2068.txt>.
52. GHOZAL Ahmed El Akhdar. Les appareils d'impression électronique. L'ordinateur et le caractère arabe : problématique et solution. In : . pp.244-354. (En langue arabe)

53. GRISLIN Martial, Christophe KOLSKI. Évaluation des interfaces Homme-Machine lors du développement des systèmes interactifs. *Technique et Science informatiques*. Vol.15-n°3/199.
54. HABERMAS J. La technique et la science comme « idéologie ». Paris : Gallimard, 1973.
55. HEBENSTREIT Jacques. Principes Informatiques : Algorithmes et décidabilité. In : *Encyclopædia Universalis France S.A.* 1995. 12-306 b.
56. HICKS N., STREETEN P. Indicators of Development : the Search for a basic needs yardstick. *World Development*. N°7, 1979 ; pp. 567-580.
57. HUDRISIER Henri. La norme de représentation des caractères sur plusieurs octets : (ISO-IEC/10646), un enjeu pour le Sommet d'Hanoi. MED'IACOM INFOROUTES : un défi pour le monde arabe, la Méditerranée et la Francophonie. Université de Paris VIII. Beyrouth les 9, 10 et 11 avril 1997
58. RAMONET Ignacio. Québec et Mondialisation. *Le Monde Diplomatique*, Avril 1996 ; p.1
59. Information et documentation – Translittération des caractères arabes en caractères latins. Partie 2 : Langue arabe – Translittération simplifiée. ISO 233-2 : 1993 (F) 6p.
60. International Commission for the Study of Communication Problems (MacBride Commission). *Many Voices. One World*. Paris : Unesco, 1980.
61. International Labor Organization. *Employment, Growth and Basic Human Needs : A One-World Problem*. Genève : BIT, 1976
62. International Standard ISO/IEC 2022 : Information technology-Character code structure and extension techniques. 4è éd.. 1994.47 p.
63. Internet Mail Consortium. Using International Characters in Internet Mail. Internet Mail Consortium Report : MAIL-I18N IMCR-010, 1 Aout, 1998. <http://www.imc.org/mail-i18n.html>
64. IRSIT, Tunis. Arabic/latin bilingual videotex, presentation syntax : Proposals for the standardization. 1991. 26 p. Draft.
65. ISO 8859. Traitement de l'information - jeux de caractères graphiques codés sur un seul octet : partie 6, alphabet latin/arabe. 1987. 5p.
66. ISO/IEC 9541-1:1991 Information Technology -- Font Information interchange -- Part 1 : Architecture.
67. JAKOBSON Roman. *Essais la linguistique Générale*. Paris : Minuit, 1963. Cité par Robert Page. Les processus de la communication. In : *Encyclopædia Universalis France S.A.* 1995. 6-198 b.
68. KHOULOUGHLI D.E. Linguistique arabe et informatique : recherches actuelles dans les universités de Paris. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. p.4.
69. KIMURA Yasuhiko. Néo-tech : le Zéro Philosophique. <http://www.neo-tech.com/french-zero/front.html>
70. KOLSKI Christophe. *Ingénierie des interfaces Homme-Machine : conception et évaluation*. Paris : Hermès, 1993.

71. KUENZ Pierre. Linguistique et littérature. In : Encyclopædia Universalis France S.A.1995. 13-857 b.
72. KWAN Stuart, GILROY James. Using the UTF-8 Character Set in the Domain Name System. Internet Draft. Microsoft Corp. Mars 1998. <http://draft-skwan-utf8-dns-01.txt>
73. La normalisation du classement alphabétique. Interface. Octobre 1997. Vol.8. n°6. URL : <http://www.crim.ca/APIIQ/interfac/>
74. LABED Lamia, SALHI Ridha, GHAZALI Salem. Génération de textes dans un système de traduction anglais-arabe : Torjmane. In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.91-114.
75. LABONTE Alain. Intervention au Forum « URL Internationalization » le mercredi 19 février 1997 09:37:07-0500. <http://www.acl.lanl.gov/URI/archive/uri-96.messages/>
76. LADRIERE Jean. Cybernétique. In : Encyclopædia Universalis France S.A.1995.6-1001 a.
77. LAZAR Judith. La science de la communication. Paris: PUF, 1992. (Que sais-je, 2634). p.18.
78. LE COADIC Yves-François. La science de l'information. Paris : PUF, 1994.-(Que Sais-je?, 2873).- p.3
79. LELOU Cathérine. Moteurs d'indexation et de recherche : environnement client-serveur, Internet et Intranet. Paris : Eyrolles, 1998. 277 p.
80. LEMPEN Blaise. La prothèse informatique. Nouvelles de la science et des technologies, vol.11, n°1, 1993 ; pp. 75-80.
81. LEVIS STRAUS C. Introduction à l'œuvre de Marcel Mauss. In : M.Mauss, Sociologie et Anthropologie, éd. C. Lévis Straus, préf. G.Gurvitch,. 9e éd. Paris : PUF, 1985.
82. LOLOU Cathérine. Moteurs d'indexation et de recherche. Op. Cit. p.43.
83. LUHMANN Niklas. Communication et action. Réseaux Communication Technologie Société. N°50, 1991 ; pp. 131-156.
84. LUNDU M.G. The justification of establishing and maintaining a nation-wide library service ; with special reference to the library and the process of national development in Zambia. Thèse de PhD, Université de Sheffield, 1982.
85. MACKENZIE Charles E. Coded Character Sets, History and Development. London : Addison-Wesley Publishing Company, 1980. p.213
86. MALABOU Catherine. DERRIDA J. In : Encyclopædia Universalis France S.A.1995.
87. MALONEY Murray. Hypertext links in HTML. 1995. <http://www.sil.org/sgml/maloneyRelationships.html>
88. MARTIN William J. The information society : idea or entity. Aslib Proceedings. Vol.40, n°11-12, novembre 1988 ; pp.303-309.

89. MASINTER Larry. Martin Düerst. Representing non-ASCII Characters in URIs and Extended URIs. Internet Draft. Août 1998.
90. MASUDA Y. C. In : The Information technology revolution. Ed. by Tom Forrester, Oxford, Blackwell, 1985 ; pp.620-634.
91. MERAYATI Mohamed. Le fonctionnement des outils et des équipements avec le caractère arabe. In : L'usage de la langue arabe en informatique. Tunis : ALECSO, 1996. pp.76-117.(EN langue arabe)
92. MERAYATI Mohamed. Le fonctionnement des outils et des équipements avec le caractère arabe. Op.Cit. p.86.
93. Microsoft MS DOS User's Guide : Arabic Supplement. Microsoft Corporation, 1988. p.18
94. MIGNOT-LEFEBVRE Yvonne. Technologies de communication et d'information : une nouvelle donne internationale ? Revue Tiers Monde, t.XXX, n°138, avril-juin 1994. p.249
95. MOCKAPETRIS P. Domain names – Concepts and facilities. Network Working Group, RFC 1034. November 1987.
96. MOCKAPETRIS P. Domain names – Implementation and specification. Network Working Group, RFC 1035. November 1987.
97. MOUNIN Georges. Linguistique : Objet et Méthodes. In : Encyclopædia Universalis France S.A.1995. 13-825 a.
98. MUSTAPHA SULEIMAN H. Sorting problems of the standard arabic character set. Computer Standards & Interfaces, 1996, n° 18, pp.159-173
99. National Center for Supercomputing Applications. Université de l'Illinois, Urbana-Champaign. <http://www.ncsa.uiuc.edu/General/NCSAHome.html>
100. NEVEU Erik, RIEFFEL Rémy. Les effets de réalité des sciences de la communication. Réseaux Communication Technologie Société. N°50, 1991 ; pp. 13-39.
101. NEY Georges. Mesures électriques et électroniques. In : Encyclopædia Universalis France S.A.1995. 15-41 c
102. NICOL Gavin T. The Multilingual World Wide Web. Electronic Book Technologies. Inc. <http://www.sil.org/sgml/nicol-multwww.html>
103. NIELSEN J., MOLICH R., Heuristic evaluation of users interfaces, Actes de colloque CHI'90, Seattle, ACM New York, pp.349-356, 1990
104. NOERR Peter L. Information Management and Engineering Ltd. London. Voir à l'adresse [http :
http://www.ua.ac.be/KB/pn/pnoerr6.html](http://www.ua.ac.be/KB/pn/pnoerr6.html)
105. NORA S., MINC A. L'informatisation de la société. Paris : La Documentation Française, 1978.
106. Norme Internationale : ISO 8859-6, Traitement de l'information - Jeux de caractères graphiques codés sur un seul octet. Première édition, 1987-08-15. p.1.
107. OCHAI Adakole. The emerging information society. International library review. N°16, 1984 ; pp.367-372.

108. Petit Robert : dictionnaire analogique et alphabétique de la langue française. Nouvelle édition, version électronique. 1996.
109. PIERRE Samuel, COUTURE Marc. Télécommunications et transmission de données. Uébec : Télé-université, 1992. p.13.
110. RACHED Roshdi, ANAWATI Georges C. Islam : les mathématiques et les autres sciences. In : Encyclopædia Universalis France S.A.1995. 12-706 b
111. REIX Robert. Traitement des données. Paris : Foucher, 1984. p.189.
112. RICOEUR Paul, SEBESTIK Jean. Philosophie du langage. In : Encyclopædia Universalis France S.A.1995. 13-436 a.
113. RICOEUR Paul. Signes et Sens. Encyclopædia Universalis, France S.A, 1995.
114. ROBBINS K. New technology : the political economy of general Ludd. In : Information technology : impact of the way of life, vol.2. Dublin : National Board for science and technology, 1981.
115. RONDEAU Jean-Claude. Multilinguisme et génie linguistique. Université, mai-juin 993 ; p.46.
116. ROQUEPLO P. Penser la technique. Paris : le Seuil, 1983 ; p.43.
117. ROUBAN Luc. Commentaire général : quelle politique pour l'information scientifique et technologique ? Nouvelles de la science et des technologies. Vol.1, n°1, 1993 ; pp.81-87.
118. ROUBINE Elie. Théorie du Signal. In : Encyclopædia Universalis France S.A.1995. 20-1071 b
119. SABAH Gérard. Groupe Langage et Cognition.
<http://www.limsi.fr/Individu/gs/GroupeLC/Theme3.html>
120. SELIGMAN Ruth. The Buenos Aires plan of action for promoting and implementing technicalcooperation among developing countries : ten years after. Cooperation south. Décembre 1977 ; p. 11
121. SENACH B., ALENGRY P. Évaluation d'un dispositif d'assistance à la conduite et au dépannage de chaînes de fabrication automatisées. rapport INRIA n°163. Mars 1994.
122. SFEZ Lucien. La communication. Paris : PUF, 1994. (Que sias-je).127 p.
123. SHANNON C.E., WEAVER W. Théorie mathématique de la communication. Paris : C.E.P.L., 1975.
124. ISO-8879. Information Processing Text and Office Systems — Standard Generalized Markup Language (SGML). 1986.
125. Social and historical aspects of mathematics. University of Wolverhampton. School of Computing and Information Technology. Module MM2217
<http://www.scit.wlv.ac.uk/university/scit/modules/mm2217/ocm.htm>
126. SURPRENANT Tom. Future libraries : the electronic environment. Wilson Library Bulletin, Janvier 1982.

127. SWEENEY M. et al. Evaluating user-computer interaction : a framework. *International Journal Man-Machine Studies*, 38, 689-711, 1993.
128. Technical cooperation conference adopts plan of action : self-reliance stressed. *UN Chronicle*, octobre 1978 ; pp. 43-46
129. The Unicode Standard, Version 2.0. Addison Wesley Longman, 1996.
130. THIMBLEY Harold. User interface Design. New York : ACM Press, 1955. pp.6-7.
131. UTF-8, a Transformation Format of Unicode and ISO 10646. <http://www.imc.org/rfc2279>
132. VAN WINGEN Johan W. Coded character sets and programming languages. *Leiden, the Netherlands* : ISO/IEC JTC1/SC2 N 1961R, ISO/IEC JTC1/SC22 N 578R, September 1988. Revised April 1989. Version 1.2
133. VANSTEEN Jean Michel. Modems-Howto. 13 février 1996. <http://www.gilde.asso.fr/miroirs/kheops/howto/Modems-HOWTO/Modems-HOWTO-1>
134. W3C activity: Internationalization. <http://www.w3.org/International/Activity.html>
135. WEAVER Warren. Contributions récentes à la théorie mathématique de la communication. In : Daniel Bounoux. *Sciences de l'information et de la communication*. Paris : Larousse, 1993. p.419.
136. Web Internationalization & Multilingualism Symposium. Seville, 20-22 November 1996. <http://www2.echo.lu/oii/en/w3c-int.html>
137. WERIBEL Stuart, GODBY Jean, Eric Miller, Ron Daniel. OCLC/NCSA Metadata Workshop Report. 1995. http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html
138. WHITEFIELD et al. A framework for human factors evaluation. *Behavior & Information Technology*, vol. 10, n°1, pp.56-79, 1991.
139. YERGAU François. Internationalization of the Hypertext Markup Language. <http://www.sil.org/sgml/rfc2070.html>
140. YERGEAU François. Internationalization of URLs. Alis Technologies. Septembre 1996. <http://ds.internic.net/rfc/rfc2044.txt> ou <http://www.alis.com:8085/~yergeau/url-00.html>

BIBLIOGRAPHIE RECOMMANDÉE

1. ABBA L. Principe pour l'établissement d'une interconnexion entre les réseaux informatiques des pays africains. Actes de la réunion de lancement du réseau d'échanges d'informations entre institutions en Afrique. Dakar, Sénégal. 27-29 février 1992.
2. ABDELAZIZ M.H. Al Mustalah Al 'Ilmi : Beina Al tarjama Wa Al Ta'rib = Le terme scientifique : entre la traduction et la translittération / Ghani ABDELAZIZ. Al Manhal, vol.55, n°504, pp.108-116, avril-mai 1993.
3. ABDUL GHANI. Arabic literature: uniterm indexing system for storage and retrieval / ABDUL GHANI. International Library Review; ISSN 0020-7837; GBR; 1987; VOL. 19; NO. 4; PP. 321-333
4. ABED, Lamia. Génération de textes arabes dans un système de traduction anglais-arabe : Tordjmane / Lamia LABED; Ridha SALHI; Salem GHAZALI .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.91-114
5. ABI-RACHED Naoum. Manuel de traduction de l'arabe technique / Naoum Abi-Rached. Paris : Maisonneuve & Larousse, 1992.-187 p.
6. ACHOUR Mohamed. Système Alpha : poste de travail de traduction / Mohamed ACHOUR. IRSIT Bulletin (Tunisie), juin 1991, pp.17-18
7. ACHRATI Souleiman. Atta'rib : Al Istratijia Wa Al Tarikh : L'arabisation : la stratégie et l'histoire. Al Manhal, vol.55, n°504, pp.96-102, avril-mai 1993
8. ACIER B. Contribution à la conception et à l'évaluation d'interfaces Homme-Machine par l'analyse du comportement visuel d'un superviseur humain. Thèse de doctorat, université de valencienne, France, janvier 1990.
9. ADAKOLE Ochai. The emerging information society. International Library Review, N° 16, 1984; pp.367-372
10. AFNOR. Ergonomie et conception du dialogue homme-ordinateur - partie 1 Concepts généraux (Z67-110).-AFNOR : Janvier 1988, 19 p.
11. AFNOR. Traitement de l'information, ergonomie et conception du dialogue homme-ordinateur, Partie 1 : concepts généraux. Norme Z 67-110. Association Française de normalisation, Paris la défense, janvier 1988.
12. AL BADR B. Survey and bibliography of Arabic optical text recognition / B.AL BADR; SA MAHMOUD . Signal processing; ISSN 0165-1684; Coden SPRODR; NLD; 1995; VOL. 41; NO. 1; PP. 49-77

13. AL-FEDAGHI Sabah. Morphological Compression of Arabic Text / Sabah AL-FEDAGHI, Humoud B. AL-SADOUN. *Information Processing & Management*, vol.26, n°2, 1990, pp. 303-316
14. AL-HAWAMDEH S. Using nearest-neighbour searching techniques to access full-text documents / S. AL-HAWAMDEH, R.DE VERE, G.SMITH, P.WILLETT. *Online Review*, 1991. vol 15, n° 3/4, pp.173-191
15. AL-KHARASHI, Ibrahim. Comparing words, stems and roots as index terms for an arabic information retrieval systems / Ibrahim A.AL KHARASHI; Marta W.EVENS .- In : *Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines*. Casablanca : 8 et 9 Décembre 1993. pp.297-310
16. Alef : Guide de l'utilisateur et manuel de référence .- Tunis: IRSIT, 1991.-22 p.
17. ALIPRAND JM. Arabgic script on RLIN / JM ALIPRAND. *Library hi tech*; ISSN 0737-8831; Coden LIHTD2; USA; 1992; VOL. 10; NO. 4; PP. 59-80; BIBL.
18. Alis taftahou Abwab « Internet » lilmoustakhdam Al arabi : « Al Moustakchif » moutasaffih lichabaka billoughati Al Arabiya = Alis ouvre les portes d' « Internet » à l'utilisateur arabe : « Le fureteur » baladeur arabisé sur le réseau. *Al-Computer, communications & Electronics*, pp.71-73, Janvier 1996
19. AMARA Fethi. ARABIZATION OF INTERNET / Fethi AMARA. *Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes*. Tunis, 6-8 juin 1995
20. AMEUR, A. L'arabe manuscrit et sa reconnaissance informatique / A. AMEU; K. ROMEO-PAKER; Y. LECOURTIER.- In : *Langue Arabe et Technologies*
21. ANANE Fethi. Perception du Vidéotex en Tunisie / Mohamed ANANE. *IRSIT Bulletin (Tunisie)*, juin 1991, pp.6-8
22. ANDLER Daniel. Introduction aux sciences cognitives / Daniel ANDLER. Paris : Gallimard, 1992.
23. AUFAURE-PORTIER M.A. A high level interface language for GIS / M.A. Aufaure-Portier -- In: *JOURNAL OF VISUAL LANGUAGES AND COMPUTING*, 6(2) 1995 p. 167-182
24. AZZEDINE Mohamed. Transparence et différentes approches de l'arabisation : arabisation et transparence / Mohamed Azzedine. *Journées informatiques et langue arabe*. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 6 p.
25. B'VITAL. L'environnement traductionnel : la station de travail du traducteur de l'an 2001. 2^e journées scientifiques, 25-27 avril 1991, Mons Belgique, Édition : Réseau thématique de recherche : lexicologie, terminologie, traduction (Montréal).
26. BALBO S. Évaluation ergonomique des interfaces utilisateurs : un pas vers l'automatisation / S. BALBOS. Thèse de doctorat, Université de Grenoble 1, 1994.

27. BALKANSKI Cécile. Intentions et croyances dans le dialogue Homme-Machine : indices sémantiques / Cécile BALKANSKI .- Techniques et sciences informatiques, vol. 15, n°2/1996, pp.201-231
28. BASTIEN C. Validation de critères ergonomiques pour l'évaluation d'interfaces utilisateurs / C. BASTIEN. Rapport de recherche n°1427 INRIA. Mai 1991.
29. BASTIEN J.MC. Ergonomic criteria for the evaluation of human computer interfaces / J.MC. BASTEIN; D.L. SCAPIN. Rapport Technique INRIA n°156, juin 1993.
30. BEAUDOUIN-LAFONT, M. Interfaces Hommes-Machines : Vue d'ensemble et perspectives / M. BEAUDOUIN-LAFFONT. Génie Logiciel & systèmes experts, n° 24, Septembre 1991, pp.4-16
31. BELAID A. Le traitement de l'écriture et des documents écrits : analyse de documents de l'image à la sémantique / A.BELAID, K.TOMBRE. La tribune des industries de la langue, n°9, 1992.
32. BELHADJ KACEM Rafik. Approche d'une solution globale arabisée / Rafik BELHADJ KACEM, Walid SAYAD. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 5 p.
33. BEN HASSINET T. Contribution à la mise en œuvre d'une méthodologie de conception des interfaces Homme-Machine / T.BEN HASSINET. Rapport de DEA, université de Valenciennes, juillet 1991
34. BEN HENDA Mokhtar. Les inforoutes, objectif ou tremplin pour le développement : lecture dans le contexte tunisien / Mokhtar BEN HENDA. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. pp.52-68
35. BENAÏSSA M.L. Conception et évaluation d'une interface Homme-Machine d'un procédé ferroviaire / M.L. BENAÏSSA, H.EZZEDINE, J.C. ANGUE. Forum SCGM Transport'92, Montréal, Canada, 1-5 juin 1992
36. BENAÏSSA, M.L. Une démarche de conception, réalisation et évaluation d'IHM : application au projet ferroviaire ASTREE / M.L BENAÏSSA,. Thèse de doctorat, université de Valenciennes, 1993.
37. BENYSH D.V. A comparative review of knowledge structure measurement techniques for interface design / D.V BENYSH. International Journal of Human-Computer Interaction, 5 (3), p.211-237, 1993
38. BEYLE Henri. La longue marche de l'OCR / Henri BEYLE. Archimag, n°42, Mars 1991, pp.26-27. L'OCR est maintenant opérationnelle. Elle est le fruit d'une longue recherche. Son histoire est jalonnée d'espoirs et de prototypes. Ses méthodes d'analyses sont actuellement diverses et éprouvées
39. BILLINGSLEY P.A. Taking panes : issues in the design of windowing systems / P.A.BILLINGSLEY. In : Handbook of human-computer interaction. M. Helander (ed.), Elsevier Science Publishers B.V., North-Holland, 1988
40. BINOT J.-L. Représentation sémantique et interprétation dans une interface en langage naturel / J.-L. BINOT, L.DEBILLE, D.SEDLOCK, B.VANDECAPELLE. pp.56-82

41. BISSERET André. Concevoir une « compréhension » Homme-Machine / André BISSET. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.11-40
42. BISSON P. Interfaces Homme-Machine multimodales et adaptatives pour systèmes opérationnels / P.BISSON, P.SOULARD, A.RAIMONDO, P.AKNIN. Conférence informatique 92 « L'interface des mondes réels et virtuels », EC2, Montpellier, 23-27 Mars 1992
43. BLANQUET Marie-France. Intelligence artificielle et système d'information : le langage naturel / Marie-France BLANQUET. Paris : ESF Editeur, 1994. 267 p. (Systèmes d'Information et Nouvelles Technologies / AUPELF-UREF)
44. BLANQUET Marie-France. L'Industrie de l'Information : l'offre et la demande / Marie-France BLANQUET. Paris : ESF Editeur, 1994. (Systèmes d'Information et Nouvelles Technologies / AUPELF-UREF)
45. BORNES Christian. Les Interfaces intelligents dans l'information scientifique et technique : cours organisé par l'INRIA du 18 au 22 mai 1992 / Christian Bornes, Dir .- Paris : INRIA, 1992 .- 190 p.-ISBN 2-7261-0726-5
46. BORREGO Jorge. Le Développement de l'Industrie au Mexique. Revue Tiers Monde, n°138, avril-juin 1994; pp.313-341.
47. BOUALEM, Abdelmalek. Système de conversion de formalisation de langages naturels dans un environnement à syntaxe contrôlé : traduction automatique multilingue / Abdelmalek BOUALEM. Thèse de doctorat , INRIA, Sophia Antipolis. 1993.
48. BOUDRAA, Malika. Élaboration d'une base de données arabe phonétiquement équilibrée / Malika BOUDRAA; Bachir BOUDRAA; Bernard GUERIN .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.172-187.
49. BOUNEMRA Karima. Normalisation et arabisation / Karima BOUNEMRA. IRSIT Bulletin (Tunisie), juin 1991, pp.14-16
50. BOUYER Gérard. Transmission et réseaux de données : introduction aux éléments fondamentaux et pratiques / Gérard Bouyer.-Paris : Dunod, 1995.-450 p
51. BRINKHOFF N. Vers des normes pour les technologies linguistiques / N.BRINKHOFF. XIII Magazine, oct., 1993.
52. BUKHALID, Nabil. Lebanese Networking Experience / Nabil Boukhalid. In : Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tunis, 6-8 juin 1995
53. BUREAU VAN DIJK. Indexation assistée par ordinateur : les clés, des ouvertures. Journée d'études du 15 février. Paris : Van Dijk, 1990.
54. CACALY Serge. Le paysage documentaire de demain ou le triomphe des interfaces / Serge CACALY. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.177-182

55. CAIL F. Présentation de l'information sur écran de visualisation, revue bibliographique / F.CAIL. cahiers de notes documentaires, INRS, n°23, 1986
56. CARRENO Orlando. Le son et l'interactivité dans les hypermédias / Orlando CARRENO. Communication et langages. N°95, 1993
57. CARTIER Michel. La navigation à l'écran / Michel CARTIER .Communication et langages. N°84, 1990. pp.98-115
58. CARTIER Michel. Un code universel pour la télématique / Michel Cartier, Roger Breton, Gérard Blanchart (pres.) . Communication et langages. N°, 1971. pp.64-69.
59. CHALOUHI Adel. IDOS : une solution optimale et unifiée pour l'arabisation transparente des systèmes informatiques / Adel Chalouhi. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 17 p.
60. CHAMSIN Radhouan. Malaf Al Ta'rib Al Computiri = Dossier de l'arabisation sur ordinateur / Radhouan CHAMSIN, Charles CHELHOUB, Jamana BAIZID, Hassan AL MORTADHI, Samar NEJM. Al-Computer, communications & Electronics, pp.71-83, Janvier 1996
61. CHARTRON Ghislaine. Nouvelles problématiques pour l'IST / Ghislaine Chartron. Documentaliste Sciences de l'Information, vol.3é, n°6, 1995, pp.289-295
62. CHAUMIER J. L'indexation assistée par ordinateur : principes et éthodes / J. CHAUMIER, M. DEJEAN. Documentaliste Sciences de l'Information, vol.29, n°1, Janvier-Février 1992.
63. CHAUMIER J. L'indexation documentaire : de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique / J. CHAUMIER, M. DEJEAN. Documentaliste Sciences de l'Information, vol.27, n°6, Nov.-Déc.1990.
64. Client/Server architecture applications up more than 60 percent; Reengineering continues to increase and pay off for big companies,- EDI Monthly Report 8, May 1995, 3-5
65. CLOUTIER Jean. L'audioscriptovisuel et le multimédia / Jean COUTIER. Communication et langages, n°99, 1994
66. COUSINS Shirley Anne. Enhancing subject access to OPACS : Controlled vocabulary Vs Natural languages / Shirley Anne COUSINS. Journal of Documentation, vol.48, n°3, Septembre 1992, pp.291-309
67. COUTAZ J. Évaluation des interfaces utilisateur : taxonomie et recommandations / J.COUTAZ, S.BALBOS. Actes des sixièmes journées sur l'ingénierie des interfaces Homme-Machine, IHM'94, 211-218, Lille, 8-9 décembre 1994.
68. COUTAZ J. Interfaces hommes-ordinateurs : conception et réalisation / J.COUTAZ. Paris : Dunod, 1990
69. COUTAZ J. Interfaces Homme-Machine : un regard critique / J.COUTAZ. Techniques et sciences informatiques, 10 (1), 1991
70. CUTLER, Matthew. The well-rounded webmaster. / Matthew Cutler, Eric Richard -- In: INTERNET WORLD, 6(10) Oct. 1995 p. 30-32 . The young WWW is changing rapidly, and the job description of a webmaster is being continually upgraded. Here's the latest release.]

71. DACHELET R. Etat de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information / R Dachelet,. Cours INRIA : le document électronique, Chatelaillon, juin 1990, pp. 107-132
72. DAHMANE Majid. Information scientifique et technique (IST). Communication et langages. N°90, 1991; pp.104-105
73. DALBIN Sylvie. Interfaces dans les systèmes d'aujourd'hui / Sylvie DALBIN. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.131-173
74. DANIELS P.J. Cognitif models in information retrieval : an evaluative review / P.J. DANIELS. Journal of Documentation, vol.42, n°4, Décembre 1986, pp.272-304. Quelques travaux récents et courant dans le domaine de la modelisation cognitive sont revus par l'auteur. Un intérêt particulier est donné aux modèles utilisateur.
75. DE ROSNEY Joël. Ce que va changer la revolution informationnelle. Le Monde Diplomatique, Août 1996; p.19
76. DENIS M. Approches cognitives de la description. In Progrès de la Recherche Cognitive / M. Denis & F.Robinson. Actes du 4^e colloque de l'Association pour la recherche Cognitive. Publication INRIA, 1990, 129 p-143.
77. DESCAMPS WASSIF S. La banque de données A'LAM de l'Institut du monde arabe / Wassif S. DESCAMPS. Documentaliste Sciences de l'Information; 1995; VOL. 32; NO. 1; PP. 39-42
78. DESSAIGNE M.F. Man-machine interface design / M.F.DESSAIGNE. 11th Congress of the International Ergonomics association. Paris, july 15-20, 1991
79. DEVEZE Jean. L'information scientifique et technique produit scientifique ou discours sur la science. Communication et langages. N°90, 1990. pp.95-23
80. DEVEZE-BERTHET Denise. La promotion de la culture scientifique et technique : enjeu politique ou facteur d'émancipation / Denise DEVEZE-BERTHET. Nouvelles de la Science et des Technologies, vol.11, n°1, 1993, pp.61-66
81. DJOUDI Mahieddine. Contribution à l'étude et à la reconnaissance automatique de la parole en arabe standard / Mahieddine DJOUDI; Jean Paul HATON (dir.) . Th. doct. : Inform./Université de Nancy 1. FRA/1991/91NAN10395. 1992-11; 150 p
82. DJOUDI, Mohieddine. Reconnaissance de la parole en arabe moderne : le système MARS / Mohieddine DJOUDI .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.135-143
83. Dominique HOELTGEN. L'Afrique sur Internet / Dominique HOELTGEN. Jeune Afrique, n°1821 du 30 Novembre 1995.pp 59-62
84. DORDICK Herbert. Information society indicators : description, measurement, prediction. In : Information societies : comparing the Japanese and American experiences. Seattle : International Communication Centre, 1978.

85. DOSA Marta. Information Transfer as Technical Assistance for Development. *Journal of the American Society for Information Science*. Vol.36, n°3, 1985; pp.146-152.
86. DUBOIS Daniel. Sémantique et cognition : catégories, prototypes, typicalité / Daniel DUBOIS. Paris : CNRS, 1991.
87. DUCASSE Roland. De la galaxie Guttemberg au Campus virtuel / Roland DUCASSE. Université, vol.15, n°3, octobre 1994
88. DUCASSE Roland. Inforoutes et infosystèmes. De l'architecture client-serveur aux systèmes d'information distribués : la culture technique des usagers en question; éléments pour une rénovation de compétences / Roland DUCASSE. Actes du séminaire international sur la psychologie et comportement des usagers face aux nouvelles technologies de l'information organisé par l'ISD et BIEF. Tunis, 16-18 novembre 1994. pp 43-45
89. DUCASSE Roland. Modèle Client-Serveur et systèmes d'information distribués : éléments pour une réingénierie des services d'information scientifique et technique / Roland DUCASSE. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. pp.133-138
90. DUCLOY Jacques. Une ingénierie de l'information adaptée à Internet / Jacques DUCLOY. Actes du colloque sur la communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater/Internet organisé par le CEM-GRESIC. Bordeaux 16-18 mars 1995. pp.124-138
91. DUFOUR Arnaud. Internet / Arnaud DUFOUR. Paris : PUF, 1995.-127 p.-(Que Sais-Je?, 3073)
92. EBOUYEA Michel. Les interfaces graphiques dans la communication Homme-Machine / Michel EBOUYEA. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.41-74
93. EFTHIMIADIS Efthimis N. Online searching aids : a review of front ends, gateways and other interfaces / Efthimis N. EFTHIMIADIS. *Journal of Documentation*, vol.46, n°3, Septembre 1990, pp.218-262
94. EL ALAOUI Faris M. Alexie sans agraphie en langue arabe : étude neurolinguistique et IRM / Faris M. EL ALAOUI; BENBELAID F; ALAOUI C; TAHIRI L; JIDDANE M; AMARTI A; CHKILI T. *Revue neurologique* : (Paris); 1994; vol. 150, n°11, pp. 771-775
95. ELLIS David. The physical and cognitive paradigms in information retrieval research / David ELLIS. *Journal of Documentation*, vol.48, n°1, Mars 1992, pp.45-64
96. ELLOUZE Nouredine. Les logiciels arabisés : un créneau pour l'industrie de demain / Nouredine ELLOUZE . *IRSIT Bulletin* (Tunisie), juin 1991, pp.4-5
97. EMBERG John. Towards a new paradigm for international cooperation : telematics and computer networks, tools for sustainable development of rural and remote regions and countries / Johan EMBERG. JE/4.6.95 .- IUT, Document interne, 1995
98. EMMERIJ L. Our Grave New World, Trends in International development and Security. *Development : seeds of Change*. N°1, 1983; pp. 5-12.
99. EUROBROKERS. L'Influence du développement des techniques sur l'avenir de la traduction automatique. Luxembourg, 1990.

100. EWING J. An experimental comparison of a mouse and arrow-jump key for an interactive encyclopedia / J.EWING, S.MEHRABANZAD, S.SHECK, D.OSTROFF, B.SHNEIDERMAN. international Journal of Man-machine studies, 24, 1986.
101. FARROW John F. A cognitive process model of document indexing / John F.FARROW. Journal of Documentation, vol.47, n°2, Juin 1991, pp.149-146
102. FLICHY Philippe. Hypertexte, une représentation de l'information / Philippe FLICHY. Archimag, n°45, juin 1991.
103. FLUHR C. Génération automatique des liens hypertextes / C.FLUHR. L'informatique documentaire : Bulletin du CID, 1989, n°35, pp.89-98
104. FLUHR Christian. Le traitement du langage naturel dans la recherche d'information documentaire / Christian FLUHR. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.103-128
105. FUCHS Cathérine. Traduction automatique / Cathérine FUCHS, Lawrence DANLOS. In : Linguistique et traitements automatiques des langues .- Paris : Hachette, 1993.-pp.193-216
106. GALACSI. Les systèmes d'information : analyse et conception / Galacsi (nom collectif de : H.Briand, J.B. Crampes, C.Ducateau, Y.Hebrail, D.Dehrin-Aime, J.Kouloumadjian, R.Sabattier);-Parsi : Dunod, 1986.280 p.
107. GARDNER Howard. Histoire de la révolution cognitive : la nouvelle science de l'esprit / Howard GARDNER. Paris : Payot, 1993.
108. GARFIELD Eugene. 2001 : an information society ? Journal of Information Science. N°1, 1979; p.210.
109. GHAZAL Ahmed Lakhdar. Transparence et differents approches de l'arabisation : introduction a la langue arabe en informatique / Ahmed Lakhdar Ghazal. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 9 p.
110. GHAZALI, S. Elements of Arabic Phonetics / Salem GHAZALI. In : Applied Arabic Linguistics and Signals & Information Processing, pp.51-58. Hemisphere publishing corporation, 1987.
111. GIFFARD Anthony C. Un pont à travers le Sud : le système pilote d'information technique (SPIT) / Thomas L. JACOBSON. Revue Tiers Monde, t.XXXV, n°138, avril-juin 1994, pp.375-389
112. GOLDWASER Daniel. La lecture optique sur le chemin de l'automatisme / Daniel GOLDWASER. Documentaliste Sciences de l'Information, vol.28, n°3, 1991, pp.154-158
113. GOODMAN S.E. Computing in the Middle East / S.E. GOODMAN, J.D. GREEN. Communication of the ACM, vol.35, n°8, Août 1992, pp.21-22
114. GRANDGUILLAUME, G. Arabisation au Machrek et au Maghreb / GRANDGUILLAUME, G. In: Revue de Presse Maghreb Moyen-Orient (DZA) 1993/09-10, Vol. 378, 16-19 (IMA : 051.7. Rev.2/RES)

115. GRANDGUILLAUME, G. L'Arabisation en question / GRANDGUILLAUME, G. -- In: Langue arabe, la parole et la plume. 1996/04-06, vol. 19. 43-45 (IMA : 051.1 QAN)
116. GRAU Brigitte. Pragmatique et dialogue Homme-Machine / Brigitte GRAU, Gérard SABAH et Anne VILNAT. TSI, 13(1) : 9-30, 1994
117. GREEN Rebecca. The profession's models of information : a cognitive linguistic analysis / Rebecca GREEN. Journal of Documentation, vol.47, n°2, Juin 1991, pp.130-148
118. GRELLIER Christian. Internet et ses concurrents / Christian GRELLIER. Communication et langages. N°105, 1995. pp.
119. GRISLIN M., KOLSKI C., ANGUE J.C. Toward an organisation of man-machine interface evaluation techniques by a usability criteria grid / M.GRISLIN, C.KOLSKI, J.C. ANGUE. XII European Annual Conference on Human Decision Making and Manual Control, Kassel, 22-24 june 1993
120. GRISLIN Martial. Évaluation des interfaces Homme-Machine lors du développement des systèmes interactifs / Martial Grislin, Christophe Kolski. Technique et Science informatiques. Vol.15-n°3/199. p.266
121. GRISLIN Martila, KOLSKI Christopher. Évaluation des interfaces Homme-Machine lors du développement des systèmes interactifs / Martila GRISLIN, Christopher KOLSKI .- Techniques et sciences informatiques, vol. 15, n°3/1996, pp.265-296
122. GSI-ERLI. La maîtrise des réalisations d'applications d'ingénierie linguistique et documentaire : plaquette d'information / GSI-ERLI. 1993.
123. GUHA, Alope. Controlling the process with distributed multimedia. / Alope Guha, Allalaghatta Pavan, Jonathan C.L. Liu, Barry A.Roberts -- In: IEEE MULTIMEDIA, 2(2) Summer 1995 p. 20-29
124. HAMMOUCHE H. De la modélisation des tâches à la spécification d'interfaces utilisateurs / H.HAMMOUCHE. Rapport INRIA n°1959, juillet 1993.
125. HANDCOCK-BEAULIEU M. Les catalogues en ligne jugés par les utilisateurs / M.HANDCOCK-BEAULIEU. Bulletin des Bibliothèques de France, 1991, 36,, n°4, pp.312-316
126. HANDCOCK-BEAULIEU M. User friendliness end human-computer interaction in online library-catalogues / M.HANDCOCK-BEAULIEU .- Program, january 1992, vol.26, n°1, pp.29-38
127. HASSANEIN, Oussama. The global information superhighway : issues for the arab region / Oussama HASSANEIN. In : Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tunsi, 6-8 juin 1995
128. Herold, Phil. Interactive design methodology/ Phil Herold, Carla Merrill. In: DR. DOBB'S JOURNAL, 20(6) June 1995 p. 40-47; 98-102 . Designing a client/server hypertext help system.
129. HICKS N. Indicators of Development : the Search for a basic needs yardstick / N. Hicks, P.Streeten. World Development. N°7, 1979; pp. 567-580.

130. HIRZALLA, Nael. A multimedia query user interface / Nael Hirzalla, Ahmed Karmouch. In : Gagnon, Francois (ed.). Congrès canadien en génie électrique et informatique 1995 : compte rendu du congrès : le génie dans la cité informatique = 1995 Canadian Conference on Electrical and Computer Engineering : conference proceedings : engineering in the global village. Piscataway : IEEE, c1995. (p. 590-593) (SIRI: CON-59094)
131. HIX D. Human-Computer interface development tools : a methodology for their evaluation / D.HIX, R.S. SHULMAN. Communications of the ACM, 34 (3), pp. 74-87, March 1991.
132. HLAL Yahia. Traitement automatique de la langue arabe : Analyse morphosyntaxique de l'arabe : outils linguistiques dans les bases de données terminologiques arabes / Yahia Hlal. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 10 p.
133. HOELTGEN Dominique. L'Afrique sur Internet. Jeune Afrique, n°1821 du 30 Novembre 1995, pp 59-62
134. IGALENS Jacques. La normalisation / Jacques Igalens, Hervé Penan.-Parsi : Prtesse Univesitaire de France, 1994.-127 p. (Que sais-je ?, 1954)
135. INSTITUT REGIONAL DES SCIENCES INFORMATIQUES ET DES TELECOMMUNICATIONS. Spécifications techniques du Videotex Bilingue Arabe-Latin / IRSIT. Tunis : IRSIT, 1990.60 p.
136. INSTITUT REGIONAL DES SCIENCES INFORMATIQUES ET DES TELECOMMUNICATIONS. Éléments de spécifications techniques du Point d'Accès Vidéotex Bilingue Arabe-Latin / IRSIT. Tunis : IRSIT, 1990.18 p.
137. INSTITUT REGIONAL DES SCIENCES INFORMATIQUES ET DES TELECOMMUNICATIONS. Rapport d'activités : présentation des projets scientifiques et techniques / IRSIT. Tunis : IRSIT, 1990.57 p.
138. International Commission for the Study of Communication Problems (MacBride Commission). Many Voices. One World. Paris : Unesco, 1980.
139. International Labor Organization. Employment, Growth and Basic Human Needs : A One-World Problem. Genève : BIT, 1976
140. JACOBSON Thomas L. Les messageries électroniques et les services dans les pays du tiers monde / Thomas L. JACOBSON. Revue Tiers Monde, t.XXXV, n°138, avril-juin 1994, pp.343-358
141. JAMBI, Kamel M. An approach for segmenting handwritten arabic words / Kamel M. JAMBI .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.233-243
142. JAVED Mostafa. Design and analysis of the human-computer interface for an image database (Database) / Mostafa, Javed. Thesis (PH.D.)--The University of Texas at Austin, 1994, 124p. - (Source : Dissertation Abstracts International. Volume: 55-10, Section: A, page: 3023.).
143. JEANNERET Yves. Les images de la science / Yves JEANNERET. Communication et langages, n°99, 1994

144. JOUET J. Les nouvelles technologies de communication : orientations de la recherche / J.Jouet, S. Coudray. Paris : Unesco, 1990. Rapport Etudes et Documents d'Information n°105.
145. KARSENTY, S. La construction d'interfaces utilisateurs / S.Karsenty, Génie Logiciel & systèmes experts, N°24, Septembre 1991, pp.18-27
146. KEYSON D.K. Designing the user interface using rapid prototyping / D.K.KEYSON, K.C.PARSON. Applied Ergonomics, 21 (3), pp.207-211, september 1990
147. KHALIL Georges. Approche socio-économique de l'arabisation de l'informatique / Georges Khalil. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 9 p.
148. KHOULOUGHLI D.E. Linguistique arabe et informatique : recherches actuelles dans les universités de Paris / D.E. Kouloughli. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 15 p.
149. KOBEISI, Majida. Conception et réalisation d'un générateur graphique des caractères arabes pour une édition et une programmation en arabe / Majida Kobeisi .-- Thèse de troisième cycle Electroniques : Université Paris XI : Paris 1992 .-- 291 p. ; 21 cm.
150. KOLSKI C. A process method for the design of « intelligent » man-machine interfaces : case study : the decisional module of imagery / C.KOLSKI, M.TENDJAOUI, P.MILLOT. International Journal of Human Factors in manufacturing, 2 (2), April 1992
151. KOLSKI C. Ingénierie des IHM, conception et évaluation / C.KOLSKI. Editions Hermès, Paris, 372 p., 1993
152. KOLSKI C. Méthodes et modèles de conception des interfaces Homme-Machine / C.KOLSKI. Mémoire d'Habilitation à Diriger des Recherches, université de Valenciennes, 1995
153. KOLSKI C., MOUSSA F. Une approche d'intégration de connaissances ergonomiques dans un atelier logiciel de création d'interfaces pour le contrôle de procédé / C.KOLSKI, MOUSSA F. Quarantièmes Journées Internationales : le Génie Logiciel et ses applications, Toulouse, 9-13 décembre 1991.
154. KOLSKI C., TENDJAOUI M., MILLOT P. An « Intelligent » interface approach / C.KOLSKI, M.TENDJAOUI, P.MILLOT. In : « ergonomics of Hybrid Automated systems II », W. Karowski and M. Rahimi (editors), Elsevier Science Publisher B.V., 1990, pp. 587-594
155. KOLSKI, Christophe. Ingénierie des interfaces Homme-Machines : conception et évaluation / Christophe Kolski -- Paris : Hermès, 1993 (BPI : Bur. 5/6 Cote : 681.1 KOL)
156. KOVAL J.L. La technique du fenêtrage comme moyen facilitant le dialogue Homme-Machine dans une expérience d'une leçon de formation assistée par ordinateur. L'homme et l'écran : aspects de l'ergonomie en informatique / J.L.KOVAL. Actes du congrès de Nivelles : l'ergonomie en informatique, Editions de l'Université de Bruxelles, Belgique, 1986
157. KURDI, Mohsen Bassam. Multifont arabic/latin optical character recognition system / Mohsen Bassam KURDY; Ammar JOUKHADAR; Ahamd WABBI .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi

- Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.245-256
158. L'hypertexte : une révolution ?. Archimag, n° , Août 1990. pp.26631
159. L'OCR : méthode et application. Archimag, n°42, mars 1991
160. LAINE-CRUZEL Sylvie. Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs / Sylvie LAINE-CRUZEL. Documentaliste Sciences de l'Information, vol.31, n°3, 1994, pp. 143-147
161. LARDY jean-Pierre. les outils de recherche d'information sur Internet : guides, listes thématiques et index / jean-Pierre LARDY. Documentaliste Sciences de l'Information, vol.33, n°1, 1996, pp.33-37. C'est pour éclairer les professionnels de la documentation que Jean pierre Lardy étudie ici les outils de recherche d'information sur Internet : guides, listes et catalogues thématiques, et index créés automatiquement.
162. LAZAR Judith. La science de la communication / Judith LAZAR.-Paris : PUF, 1985.-125 p.- (Que Sais-je?, 2634)
163. LE BIHAN Jean. Les SGBD de 2^e génération : avantages et limites / Jean LE BIHAN. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.77-86
164. LE COADIC Yves François. La science de l'information / Yves François LE COADIC . Paris : PUF, 1994. 127 p. (Que Sais-Je ?, 2873)
165. LE CROSNIER H. Une introduction à l'hypertexte / H.LE CROSNIER . Bulletin des Bibliothèques de France, 1991, t.36, n°4, pp.281-294
166. LE NY, J.F. Sciences cognitive et compréhension du langage / J.F.Le Ny. PUF, 1989, 249 pp.
167. LEFEBVRE A. L'architecture client-serveur, aspects techniques, enjeux stratégiques .- Paris : Armand Colin, 1994
168. LEMPEN Blaise. La prothèse informatique / Blaise LEMPEN. Nouvelles de la Science et des Technologies, vol.11, n°1, 1993, pp.75-80
169. Les sciences cognitives. Le Courrier du CNRS. Dossiers scientifiques, 79, octobre 1992.
170. LIBOIS Louis-Joseph. Les télécommunications : technologies, réseaux, services / Louis-Joseph LIBOIS. Paris : ESF Editeur, 1994. 197 p. (Systèmes d'Information et Nouvelles Technologies / AUPELF-UREF)
171. Lisibilité et écriture télématique / par le groupe de recherche du CFPJ. Communication et langages. N°56, 1983. pp.64-83
172. LOGAN E. Cognitive styles and online behavior / E.LOGAN.- Information Processing & Management, 1991, vol.26, n°4.-pp.503-510
173. LUBKOV M. Carat, OCR, et texte intégral / M. LUBKOV. Archimag hors série 7, avril 1993.

174. LUNDU M.G. The justification of establishing and maintaining a nation-wide library service; with special reference to the library and the process of national development in Zambia. Thèse de PhD, Université de Sheffield, 1982.
175. MacKenzie, Scott . Movement time prediction in human-computer interfaces / Scott MacKenzie -- In: Baecker, Ronald M., Jonathan Grudin et al. Readings in human-computer interaction : toward the year 2000. San Francisco : Morgan Kaufman, c1995. (p. 483-493)
176. MAILLET J.-P. Les techniques de reconnaissance de caractère / J.-P. MAILLET. Archimag, n°66, août 1993.
177. Martel, Pierre. Oui au français québécois standard / Pierre Martel, Helene Cajolet-Laganiere -- In: INTERFACE, 16(5) sept./oct. 1995 p. 14-25
178. MARTIN William J. The information society : idea or entity. Aslib Proceedings. Vol.40, n°11-12, novembre 1988; pp.303-309.
179. MASUDA Y.C. The Information technology revolution. Ed. by Tom Forrester, Oxford, Blackwell, 1985; pp.620-634.
180. MENADIER, Jean Pierre. L'Interface utilisateur : pour une informatique plus conviviale / Jean Pierre Menadier. -- Paris : Dunod, 1991. (BPI : Bur. 5/6 Cote : 682 MEI)
181. MERY N. Les logiciels d'OCR / N. MERY . Archimag, n°58, octobre 1992.
182. MIGNOT-LEFEBVRE Yvonne. Technologies de communication et d'information : une nouvelle donne internationale / Yvonne Mignot-Lefebvre. Revue Tiers Monde, t.XXXV, n°138, avril-juin 1994, pp.245-277
183. MILOUDI, M.T. Approche système expert pour la traduction automatique réversible du Braille intégral ou abrégé vers l'arabe / M.T. MILOUDI; Z. ZMERLI .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.371-381.
184. MNIF, Mohamed. Arabic/Latin Bilingual Videotex System / Mohamed MNIF. Actes de la conférence régionale sur la communication et les réseaux d'information dans les pays arabes. Tounsi, 6-8 juin 1995.
185. MONEIMNE Walid. AO vers l'arabe : spécification d'une génération standard de l'arabe : réalisation d'un prototype anglais-arabe à partir d'un analyseur existant / Walid MONEIMNE; Christian BOITET (Dir. thèse); Th. doct. : Inform./Grenoble 1/1989/89 GRE1 0061; 1989; 259 p.; 30 cm; BIBL. 80 ref.;
186. MOUSSA F. Etude de l'intégration d'un outil expert interactif ergonomique dans la démarche de conception d'interface graphique Homme-Machine / F.MOUSSA, C.KOLSKI, P.MILLOT. II Colloque, « L'ordinateur, l'Homme et l'Organisation », Nivelles, 9-11, 1990
187. MUSA F.A. A system for processing bilingual Arabic/English text / F.A. MUSA . Journal of the American Society for Information Science; ISSN 0002-8231; USA; 1986; VOL. 37; NO. 5; PP. 288-293; BIBL.

188. MYERS, Brad A. State of the art in user interface software tools / Brad A. Myers -- In: Baecker, Ronald M., Jonathan Grudin et al. Readings in human-computer interaction : toward the year 2000. San Francisco : Morgan Kaufman, c1995. (p. 323-343)
189. NAZIM Ali S. CD-ROM databases as an alternative means to online information: the experience of a university library in developing countries / Ali S. NAZIM. International online information meeting. 12/1988-12-06/London (Londres); GBR; Oxford: Learned Information Europe; DA. 1988; VOL. 2; PP. 569-573; BIBL
190. NGYEN, Cao Lieu. Modelisation des donnees multimedia reparties dans des reseaux client/serveur / Cao Lieu Nguyen, Soudougou M. Konate -- In: Gagnon, Francois (ed.). Congres canadien en genie electrique et informatique 1995 : compte rendu du congres : le genie dans la cite informatique = 1995 Canadian Conference on Electrical and Computer Engineering : conference proceedings : engineering in the global village. Piscataway : IEEE, c1995. (p. 1026-1029) (SIRI: CON-59094)
191. NIELSEN J. Designing user interfaces for international use / J.NIELSEN. North-Holland, Elsevier, Amsterdam, 1990
192. NIELSEN J. Heuristic evaluation of users interfaces / J. Nielsen, R. Molich. Actes de colloque CHI'90, Seattle, ACM New York, pp.349-356, 1990
193. NIELSEN J. Hypertext and hypermedia / J.NIELSEN. Academic Press, 1990
194. NORA S., MINC A. L'informatisation de la société. Paris : La Documentation Française, 1978.
195. NOUH A. Recognition of constrained handprinted Arabic Character / A.Nouh .- In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.201-214.
196. Nouvelle technologies et communication de l'information : de l'analyse des besoins à l'ingénierie documentaire. Paris : ADBS Édition, 1994. 274 p.
197. OAKLEY B. La traduction automatique / B. OAKLEY. XIII Magazine, n°6, mai 1992.
198. OCHAI Adakole. The emerging information society. International library review. N°16, 1984; pp.367-372.
199. PAOLI Clément. Les passerelles de communication / Paoli CLEMENT. In : Christian Bornes (dir.), Les Interfaces Intelligents dans l'Information Scientifique et Technique. Cours INRIA, 1992. pp.89-101
200. PERRIAULT I. Les systèmes experts documentaires / I. PERRIAULT. Archimag, n°27, 1989.
201. PIERRE Samuel. Télécommunications et transmission de données / Samuel PIERRE, Marc COUTURE .- Québec / Télé-université, 1992. 458 p
202. PLATT Charles. Interactive Entertainment: Who Writes It? Who Reads It ? Who Needs It ? / Charles Platt .- Wired 3(9) (September 1995): 145-149, 195-197
203. POLITY Yolla. Évaluation des modes de recherche en langage naturel / Yolla POLITY. Documentaliste Sciences de l'Information, vol.31, n°3, 1994, pp.136-142

204. POLITY Yolla. Vers une ergonomie linguistique / Yolla POLITY. Documentaliste Sciences de l'Information, vol.31, n°3, 1994, p.135
205. POLLIER A. Évaluation d'une interface par des ergonomes : diagnostics et stratégies / A.POLLIER. Rapport de recherche n°1391 INRIA, février 1991 (Le Travail Humain, tome 55,n°1/1992, p.71-96)
206. PUJOLLE Guy. Les réseaux / Guy PUJOLLE.- Paris : Eyrolles, 1996.-858 p
207. Quelles langues pour la science ? / Sous la direction de Bernard CASSEN. Paris : La Découverte, 1990.
208. RAHMAN, Halima M.A. Institutions of Higher Education : progress through arabization / Halima M.A. RAHMAN -- In: Soudanaow (SDN) 1993/08, 19-20 (IMA : 051.1 SUD)
209. RAJOUANI, A. Synthèse de l'arabe à partir du texte : développement et réalisation d'un système de haute qualité / A. Rajouani; M. OUadou; M. Najim; M. Zyoute . In : Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Études Islamiques et les Sciences Humaines. Casablanca : 8 et 9 Décembre 1993. pp.117-133.
210. RAMONET Ignacio. Québec et Mondialisation. Le Monde Diplomatique, Avril 1996; p.1
211. RASTIER François. Sémantique et recherche cognitives / François RASTIER. Paris : PUF, 1991.
212. RENAUD P.E. Introduction aux systèmes client-serveur / P.E. RENAUD .- Paris : Hermès, 1994
213. RICHARD J.F. Traité de psychologie Cognitive, le traitement de l'information symbolique / J.F.RICHARD, C.BONNET, R.GUIGLIONE (Eds). Paris : Dunod, 1990
214. RICHERT, Nichole. Arabisation et Technologie / Nicholle Richert. -- In: Rabat : Institut d'tudes et de Recherches pour l'arabisation, 1987 -- 532 p.; 25 cm. (IMA : 449 RIC)
215. ROBBINS K. New technology : the political economy of general Ludd. In : Information technology : impact of the way of life, vol.2. Dublin : National Board for science and technology, 1981.
216. ROBERT J.M. Élaboration d'une grille d'analyse des recommandations ergonomiques pour la conception d'interfaces homme-ordinateur / J.M.ROBERT, C.CHAPDELAINE. Actes du XXV^e congrès de la SELF, Montréal, 3-5 Octobre 1990
217. ROLLING L. Les ressources langagières : un outil prodigieux pour la connaissance / L. ROLLING. XIII Magazine, n°9, 1993.
218. ROMAN André. Reconnaissance des schémas primitifs de la langue arabe : en vue de son traitement automatique / André Khoulooghli. Journées informatiques et langue arabe. 30 novembre - 8 décembre 1987. Institut du Monde arabe. Paris. 11 p.
219. ROUAULT Jacques. Linguistique automatique, applications documentaires / Jacques ROUAULT. Berne : Peter Lang, 1987.

- 220. ROUDAUD B. La traduction automatique : l'ordinateur au service de la traduction / B. ROUDAUD. META : Études et recherches en traductique, vol.37, n°4,1992.
- 221. ROUTT, Thomas J. Microsoft's SNA server vs. Novell's NetWare for SAA. [The move to client-server requires new ways to reach SNA hosts and data. Gateways are coming in vogue : a close look at two top contenders.] / Thomas J. Routt, Deric Villanueva -- In: BUSINESS COMMUNICATIONS REVIEW, 25(6) June 1995 p. 57-64
- 222. RUFFER, S.M. Client/server computing technology : a framework for feasibility analysis and implementation / S.M Ruffer, D. Yen, S. Lee -- In: INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT, 15(2) Apr. 1995 p. 135-150
- 223. SAAD Haj Bakry. Towards a standard Arabic information processing vocabulary / Haj Bakry SAAD. Computer standards & interfaces; ISSN 0920-5489; Coden CSTIEZ; NLD; 1988; VOL. 7; NO. 4; PP. 343-348; BIBL. 27 ref. Il est considéré que le développement d'un tel vocabulaire ne doit pas se limiter à la traduction des termes vers la langue arabe
- 224. SABAH, Gérard. CAMEL : un système multi-expert pour le traitement automatique des langues, Modèles linguistiques, 12 Fasc 1, p.95-118, 1990.
- 225. SABER Mohieddine . Al Ta'rib Wa Al Mou'asara Al Teknouloujia = L'arabisation et la contemporanéité technologique / Gérard SABAH. Al Manhal, vol.55, n°504, pp.116-125, avril-mai 1993
- 226. SANDOVAL V. SGML : un outil pour la gestion électronique de documents .- Paris : Hermès, 1994
- 227. SEBILLOTTE S. Méthodologie pratique d'analyse de la tâche en vue de l'extraction de caractéristiques pertinentes pour la conception d'interface / S.SEBILLOTTE. Rapport INRIA n°163, mars 1994
- 228. SEGURET-GUEYE M. Topographie, ou les codages visuels de l'information sur l'écran / M.SEGURET-GUEYE, C.ARGOUUD. Les Éditions d'Organisation, Paris, 1989
- 229. SELIGMAN Ruth. The Buenos Aires plan of action for promoting and implementing technical cooperation among developing countries : ten years after. Cooperation south. Décembre 1977; p. 11
- 230. SENACH B. Évaluation de l'ergonomie des interfaces Homme-Machine : du prototypage aux systèmes experts / B.SENACH. Actes du Congrès ERGO-IA '90 : Ergonomie et Informatique avancée, Biarritz, Septembre 1990
- 231. SENACH B. Évaluation ergonomique des interfaces Homme-Machine : une revue de la littérature / B.SENACH. Rapport de recherche, INRIA Sophia Antipolis, n°1180, Mars 1990, 70 p.
- 232. SENACH B. Évaluation d'un dispositif d'assistance à la conduite et au dépannage de chaînes de fabrication automatisées / B.Senach, P Alengry. Rapport INRIA n°163. Mars 1994.
- 233. SERRES Alexandre. Hypertexte : une histoire à revisiter / Alexandre SERRES. Documentaliste Sciences de l'Information, vol.32, n°2, 1995, pp.71-83.

234. SFEZ Lucien. La communication / Lucien Sfez.-Paris : Presses Universitaires de France, 1994.-127 p.-(Que Sais-je ?, 2567)
235. SHAHLA YAGHMAI. Arab-islamic cultures and one-line bibliographical systems / Yaghmai N. SHAHLA; VP DIODATO; JA MAXIN . International Library Review; ISSN 0020-7837; GBR; DA. 1986; VOL. 18; NO. 1; PP. 15-24;
236. SHAWKY Salem. The Arabic Litterature of Library and Information Science / Salem SHAWKY.J. Inf. Sci.; NLD; 1979; VOL. 1; NO 4; PP. 231-234; BIBL. 5 REF
237. SINNO-RONYS S. Les hypercatalogues - nouvelles perspectives pour les OPAC / S.SINNO-RONYS. Bulletin des Bibliothèques de France, 1991, t.36 n°4, pp.303-311
238. SURPRENANT Tom. Future libraries : the electronic environment. Wilson Library Bulletin, janvier 1982.
239. SUTTER Eric. Le marketing des services d'information : pour un usage de l'information documentaire / Eric SUTTER. Paris : ESF Éditeur, 1994. 207 p. (Système d'Information et Nouvelles Technologies)
240. SWEENEY M. et al. Evaluating user-computer interaction : a framework / M.SWEENEY. International Journal Man-Machine Studies. N° 38, 1993; pp. 689-711.
241. TAGUE J., SCHULTZ R. Some measures and procedures for evaluation of the user interface in an information retrieval system / J.TAGUE, R.SCHULTZ. 11th International Conference on Research and Development in Information Retrieval, June 13-15 1988, Grenoble, France.-pp371-386
242. TARBY J.C. Gestion automatique du dialogue Homme-Machine à partir de spécifications conceptuelles / J.C.TARBY. Thèse de l'Université de Toulouse1, 1993.
243. Technical cooperation conference adopts plan of action : self-reliance stressed. UN Chronicle, octobre 1978; pp. 43-46
244. TUDESQ André-Jean. Nouvelles Technologies de la Communication et Dépendance renforcée de l'Afrique Noire / André-Jean TUDESQ. Monde en Développement. 19(73). 1991
245. VICKERY Brian. An application of language processing for a search interface / Brian VICKERY, Alina VICKERY. Journal of Documentation, vol.48, n°3, Septembre 1992, pp.255-275.
246. VICKERY Brian. Online search interface design / Brian VICKERY, Alina VICKERY. Journal of Documentation, vol.49, n°2, Juin 1993, pp.103-187
247. VIGNAUX Georges. Les sciences cognitives : une introduction / Georges VIGNAUX. Paris : La Découverte, 1992.
248. WHITEFIELD A. A framework for human factors evaluation / A. Whitefield et al. Behavior & Information Technology, vol. 10, n°1, pp.56-79, 1991.
249. ZARKEN M.A. Allisaniat Wa Barmajatou Al loughati Alarabiyati fi Al Hassoub : La linguistique et la programmation de la langue arabe sur ordinateur / M.A.ZARKEN. Al Manhal, vol.55, n°504, pp.182-189, avril-mai 1993.

LES ANNEXES

ANNEXE A

Jeux de caractères et codage de l'information

Y compris codage de l'information audio, vidéo, multimédia ou hypermédia, protection des données dans les TI, cryptographie, codes à barres, etc.

1. ISO/IEC 646:1991 Technologies de l'information -- Jeu ISO de caractères codés à 7 éléments pour l'échange d'informations (Publiée actuellement en anglais seulement)
2. ISO 962:1974 Traitement de l'information -- Matérialisation du jeu de caractères codés à 7 éléments et de ses extensions à 7 et 8 éléments sur bande magnétique à 9 pistes de 12,7 mm (0,5 in) de large
3. ISO 1073-1:1976 Jeux alphanumériques de caractères pour la reconnaissance optique -- Partie 1: Jeu de caractères ROC-A -- Formes et cotes de l'image imprimée
4. ISO 1073-2:1976 Jeux alphanumériques de caractères pour la reconnaissance optique -- Partie 2: Jeu de caractères ROC-B -- Formes et cotes de l'image imprimée
5. ISO 1113:1979 Traitement de l'information -- Représentation sur bande perforée du jeu de caractères codés à 7 éléments
6. ISO 1831:1980 Spécifications d'impression des caractères pour reconnaissance optique
7. ISO/IEC 2022:1994 Technologies de l'information -- Structure de code de caractères et techniques d'extension (Publiée actuellement en anglais seulement)
8. ISO 2033:1983 Traitement de l'information -- Codage des jeux de caractères pour reconnaissance automatisée (MICR et ROC)
9. ISO 2047:1975 Traitement de l'information -- Représentation graphique des caractères de commande du jeu de caractères codés à 7 éléments
10. ISO 2375:1985 Traitement de l'information -- Procédure pour l'enregistrement des séquences d'échappement
11. ISO 2955:1983 Traitement de l'information -- Représentation des unités du Système international et d'autres unités dans des systèmes comprenant des jeux de caractères limités
12. ISO 3275:1974 Traitement de l'information -- Matérialisation du jeu de caractères codés à 7 éléments et de ses extensions à 7 et 8 éléments pour l'échange d'information sur cassette de bande magnétique de 3,81 mm de large
13. ISO/IEC 4873:1991 Technologies de l'information -- Code ISO à 8 éléments pour l'échange d'informations -- Structure et règles de matérialisation (Publiée actuellement en anglais seulement)

14. ISO 5218:1977 Échange d'information -- Représentation des sexes humains
15. ISO 5426:1983 Extension du jeu de caractères latins codés employé pour l'échange d'information bibliographique
16. ISO 5426-2:1996 Information et documentation -- Extension du jeu de caractères latins codés pour les échanges d'informations bibliographiques -- Partie 2: Caractères latins utilisés dans les langues européennes peu répandues et typographie obsolète (Publiée actuellement en anglais seulement)
17. ISO 5427:1984 Extension du jeu de caractères cyrilliques codés employé pour l'échange d'information bibliographique
18. ISO 5428:1984 Jeu de caractères grecs codés employé pour l'échange d'information bibliographique
19. ISO 6093:1985 Traitement de l'information -- Représentation des valeurs numériques dans les chaînes de caractères pour l'échange d'information
20. ISO/IEC 6429:1992 Technologies de l'information -- Fonctions de commande pour les jeux de caractères codés (Publiée actuellement en anglais seulement)
21. ISO 6438:1983 Documentation -- Jeu de caractères africains codés pour l'échange d'informations bibliographiques
22. ISO 6523:1984 Échange de données -- Structures pour l'identification des organisations
23. ISO/IEC DIS 6523-1 Technologies de l'information -- Structure pour l'identification des organisations et des parties d'organisations -- Partie 1: Identification des systèmes d'identification d'organisations (Révision de l'ISO 6523:1984)
24. ISO/IEC DIS 6523-2 Technologies de l'information -- Structure pour l'identification des organisations et des parties d'organisations -- Partie 2: Enregistrement des systèmes d'identification d'organisations (Révision de l'ISO 6523:1984)
25. ISO 6586:1980 Traitement de l'information -- Matérialisation des jeux de caractères codés à 7 éléments et à 8 éléments sur cartes perforées
26. ISO 6630:1986 Documentation -- Caractères de commande bibliographiques
27. ISO 6709:1983 Représentation normalisée des latitude, longitude et altitude pour la localisation des points géographiques
28. ISO 6861:1996 Information et documentation -- Jeu de caractères codés de l'alphabet glagolitique pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
29. ISO 6862:1996 Information et documentation -- Jeu de caractères codés mathématiques pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
30. ISO 6936:1988 Traitement de l'information -- Conversion entre les jeux de caractères codés de l'ISO 646 et l'ISO 6937-2 et l'alphabet télégraphique international CCITT no 2 (ITA 2)
31. ISO/IEC 6937:1994 Technologies de l'information -- Jeu de caractères graphiques codés pour la transmission de texte -- Alphabet latin (Publiée actuellement en anglais seulement)
32. ISO 7064:1983 Traitement des données -- Systèmes de caractères de contrôle

33. ISO/IEC 7350:1991 Technologies de l'information -- Enregistrement des répertoires de caractères graphiques de l'ISO/CEI 10367 (Publiée actuellement en anglais seulement)
34. ISO/IEC 7826-1:1994 Technologies de l'information -- Structure générale pour l'échange de valeurs de code -- Partie 1: Identification des systèmes de codification
35. ISO/IEC 7826-2:1994 Technologies de l'information -- Structure générale pour l'échange de valeurs de code -- Partie 2: Enregistrement des systèmes de codification
36. ISO 8372:1987 Traitement de l'information -- Modes opératoires d'un algorithme de chiffrement par blocs de 64 bits
37. ISO 8859-1:1987 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 1 : Alphabet latin no. 1
38. ISO/IEC DIS 8859-1 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 1: Alphabet latin no 1 (Révision de l'ISO 8859-1:1987) (DIS distribué en version anglaise seulement)
39. ISO 8859-2:1987 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 2: Alphabet latin no. 2
40. ISO/IEC DIS 8859-2 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 2: Alphabet latin no 2 (Révision de l'ISO 8859-2:1987) (DIS distribué en version anglaise seulement)
41. ISO 8859-3:1988 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 3: Alphabet latin No 3
42. ISO/IEC DIS 8859-3 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 3: Alphabet latin no 3 (Révision de l'ISO 8859-3:1988) (DIS distribué en version anglaise seulement)
43. ISO 8859-4:1988 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 4: Alphabet latin No 4
44. ISO/IEC DIS 8859-4 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 4: Alphabet latin no 4 (Révision de l'ISO 8859-4:1988) (DIS distribué en version anglaise seulement)
45. ISO/IEC 8859-5:1988 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 5: Alphabet latin/cyrillique
46. ISO/IEC DIS 8859-5 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 5: Alphabet latin/cyrillique (Révision de l'ISO/CEI 8859-5:1988) (DIS distribué en version anglaise seulement)
47. ISO 8859-6:1987 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 6: Alphabet latin/arabe
48. ISO/IEC DIS 8859-6 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 6: Alphabet latin/arabe (Révision de l'ISO 8859-6:1987) (DIS distribué en version anglaise seulement)
49. ISO 8859-7:1987 (Publiée actuellement en anglais seulement)

50. ISO 8859-8:1988 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 8: Alphabet Latin/Hébreu (Publiée actuellement en anglais seulement)
51. ISO/IEC DIS 8859-8 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 8: Alphabet latin/hébreu (Révision de l'ISO 8859-8:1988)
52. ISO/IEC 8859-9:1989 Traitement de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 9: Alphabet Latin no 5 (Publiée actuellement en anglais seulement)
53. ISO/IEC DIS 8859-9 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 9: Alphabet latin no 5 (Révision de l'ISO/CEI 8859-9:1989) (DIS distribué en version anglaise seulement)
54. ISO/IEC 8859-10:1992 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 10: Alphabet latin no 6 (Publiée actuellement en anglais seulement)
55. ISO/IEC DIS 8859-10 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 10: Alphabet latin no 6 (Révision de l'ISO/CEI 8859-10:1992) (DIS distribué en version anglaise seulement)
56. ISO/IEC DIS 8859-13 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 13: Alphabet latin no 7 (DIS distribué en version anglaise seulement)
57. ISO/IEC DIS 8859-14 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 14: Alphabet latin no 8 (Celte)
58. ISO/IEC DIS 8859-15 Technologies de l'information -- Jeux de caractères graphiques codés sur un seul octet -- Partie 13: Alphabet latin no 0
59. ISO 8957:1996 Information et documentation -- Jeux de caractères codés de l'alphabet hébreu pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
60. ISO 9036:1987 Traitement de l'information -- Jeu de caractères arabes codés à 7 éléments pour l'échange d'information
61. ISO 9160:1988 Traitement de l'information -- Chiffrement de données -- Conditions pour interopérabilité dans la couche physique
62. ISO/IEC 9281-1:1990 Technologies de l'information -- Méthodes de codage d'image -- Partie 1: Identification (Publiée actuellement en anglais seulement)
63. ISO/IEC 9281-2:1990 Technologies de l'information -- Méthodes de codage d'image -- Partie 2: Procédure pour l'enregistrement (Publiée actuellement en anglais seulement)
64. ISO/IEC 9282-1:1988 Traitement de l'information -- Représentation codée de l'image -- Partie 1: Principes de codage pour la représentation d'image dans un environnement codé à 7 et à 8 éléments
65. ISO/IEC 9282-2:1992 Traitement de l'information -- Représentation codée de l'image -- Partie 2: Codage en accroissement des listes de points dans un environnement à 7 ou à 8 éléments (Publiée actuellement en anglais seulement)
66. ISO/IEC TR 9789:1994 Technologies de l'information -- Principes directeurs pour l'organisation et la représentation des éléments de données -- Méthodes et principes de codage (Publié actuellement en anglais seulement)

67. ISO/IEC 9796:1991 Technologies de l'information -- Techniques de sécurité -- Schéma de signature numérique rétablissant le message
68. ISO/IEC 9796-2:1997 Technologies de l'information -- Techniques de sécurité -- Schémas de signature numérique rétablissant le message -- Partie 2: Mécanismes utilisant une fonction de hachage
69. ISO/IEC 9797:1994 Technologies de l'information -- Techniques de sécurité -- Mécanisme d'intégrité des données utilisant une fonction de contrôle cryptographique employant un algorithme de chiffrement par bloc (Publiée actuellement en anglais seulement)
70. ISO/IEC 9798-1:1997 Technologies de l'information -- Techniques de sécurité -- Authentification d'entité -- Partie 1: Généralités (Publiée actuellement en anglais seulement)
71. ISO/IEC 9798-2:1994 Technologies de l'information -- Techniques de sécurité -- Authentification d'entité -- Partie 2: Mécanismes utilisant des algorithmes de chiffrement symétriques (Publiée actuellement en anglais seulement)
72. ISO/IEC 9798-3:1993 Technologies de l'information -- Techniques de sécurité -- Mécanismes d'authentification d'entité -- Partie 3: Authentification d'entité utilisant un algorithme à clé publique (Publiée actuellement en anglais seulement)
73. ISO/IEC DIS 9798-3 Technologies de l'information -- Techniques de sécurité -- Authentification d'entité -- Partie 3: Mécanismes utilisant des techniques de signature asymétriques
74. ISO/IEC 9798-4:1995 Technologies de l'information -- Techniques de sécurité -- Mécanismes d'authentification d'entité -- Partie 4: Mécanismes utilisant une fonction cryptographique de vérification (Publiée actuellement en anglais seulement)
75. ISO/IEC DIS 9798-5 Technologies de l'information -- Techniques de sécurité -- Authentification d'entité -- Partie 5: Mécanismes utilisant les techniques de connaissance du zéro (DIS distribué en version anglaise seulement)
76. ISO/IEC 9979:1991 Techniques cryptographiques -- Procédures pour l'enregistrement des algorithmes cryptographiques
77. ISO/IEC DIS 9979 Technologies de l'information -- Techniques de sécurité -- Procédures d'enregistrement des algorithmes cryptographiques (Révision de l'ISO/CEI 9979:1991) (DIS distribué en version anglaise seulement)
78. ISO/IEC 10116:1997 Technologies de l'information -- Techniques de sécurité -- Modes opératoires d'un chiffrement par blocs de n-bits (Publiée actuellement en anglais seulement)
79. ISO/IEC 10118-1:1994 Technologies de l'information -- Techniques de sécurité -- Fonctions de brouillage -- Partie 1: Généralités (Publiée actuellement en anglais seulement)
80. ISO/IEC 10118-2:1994 Technologies de l'information -- Techniques de sécurité -- Fonctions de brouillage -- Partie 2: Fonctions de brouillage utilisant un algorithme de chiffrement par blocs de n bits (Publiée actuellement en anglais seulement)
81. ISO/IEC DIS 10118-3 Technologies de l'information -- Techniques de sécurité -- Fonctions de brouillage -- Partie 3: Fonctions de hachage dédiées (DIS distribué en version anglaise seulement)

82. ISO/IEC DIS 10118-4 Technologies de l'information -- Techniques de sécurité -- Fonctions de brouillage -- Partie 4: Fonctions de hachage utilisant l'arithmétique modulaire (DIS distribué en version anglaise seulement)
83. ISO/IEC 10367:1991 Technologies de l'information -- Jeux de caractères graphiques codés normalisés à utiliser dans les codes à un octet (Publiée actuellement en anglais seulement)
84. ISO/IEC 10538:1991 Technologies de l'information -- Fonctions de commande pour la communication de texte (Publiée actuellement en anglais seulement)
85. ISO 10585:1996 Information et documentation -- Jeu de caractères codés de l'alphabet arménien pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
86. ISO 10586:1996 Information et documentation -- Jeu de caractères codés de l'alphabet géorgien pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
87. ISO/IEC 10646-1:1993 Technologies de l'information -- Jeu universel de caractères codés à plusieurs octets -- Partie 1: Architecture et table multilingue (Publiée actuellement en anglais seulement)
88. ISO 10754:1996 Information et documentation -- Extension du jeu de caractères codés de l'alphabet cyrillique pour les langues non-slaves pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
89. ISO/IEC 10918-1:1994 Technologies de l'information -- Compression numérique et codage des images fixes de nature photographique: Prescriptions et lignes directrices
90. ISO/IEC 10918-2:1995 Technologies de l'information -- Compression et codage numériques des images fixes à modelé continu: Tests de conformité
91. ISO/IEC 10918-3:1997 Technologies de l'information -- Compression et codage numériques des images fixes à modelé continu: Extensions
92. ISO/IEC DIS 10918-4 Technologies de l'information -- Compression numérique et codage des images fixes de nature photographique: Enregistrement des profils JPEG, profils SPIFF, "SPIFF tags", espaces de couleur SPIFF, marqueurs APPN, types de compression SPIFF et autorités d'enregistrement (REGAUT)
93. ISO/IEC 11172-1:1993 Technologies de l'information -- Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s -- Partie 1: Systèmes
94. ISO/IEC 11172-2:1993 Technologies de l'information -- Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s -- Partie 2: Vidéo
95. ISO/IEC 11172-3:1993 Technologies de l'information -- Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s -- Partie 3: Audio
96. ISO/IEC 11172-4:1995 Technologies de l'information -- Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s -- Partie 4: Essais de conformité (Publiée actuellement en anglais seulement)
97. ISO/IEC DTR 11172-5 Technologies de l'information -- Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s -- Partie 5
98. ISO/IEC 11179-3:1994 Technologies de l'information -- Spécifications et normalisation des éléments de données -- Partie 3: Attributs de base des éléments de données

99. ISO/IEC 11179-4:1995 Technologies de l'information -- Spécifications et normalisation des éléments de données -- Partie 4: Règles et directives pour la formulation des définitions des données
100. ISO/IEC 11179-5:1995 Technologies de l'information -- Spécifications et normalisation des données -- Partie 5: Principes de dénomination et d'identification des données (Publiée actuellement en anglais seulement)
101. ISO/IEC 11179-6:1997 Technologies de l'information -- Spécifications et normalisation des éléments de données -- Partie 6: Enregistrement des éléments de données (Publiée actuellement en anglais seulement)
102. ISO/IEC 11544:1993 Technologies de l'information -- Représentation codée des images et du son - - Compression progressive des images en deux tons
103. ISO/IEC 11770-1:1996 Technologies de l'information -- Techniques de sécurité -- Partie 1: Cadre général (Publiée actuellement en anglais seulement)
104. ISO/IEC 11770-2:1996 Technologies de l'information -- Techniques de sécurité -- Gestion de clés - - Partie 2: Mécanismes utilisant des techniques symétriques (Publiée actuellement en anglais seulement)
105. ISO/IEC DIS 11770-3 Technologies de l'information -- Techniques de sécurité -- Gestion de clés -- Partie 3: Mécanismes utilisant des techniques asymétriques (DIS distribué en version anglaise seulement)
106. ISO 11822:1996 Information et documentation -- Extension du jeu de caractères codés de l'alphabet arabe pour les échanges d'informations bibliographiques (Publiée actuellement en anglais seulement)
107. ISO/DIS 13230 Optique ophtalmique -- Spécification des codes barres
108. ISO/IEC TR 13335-1:1996 Technologies de l'information -- Lignes directrices pour la gestion de la sécurité des technologies de l'information (TI) -- Partie 1: Concepts et modèles pour la sécurité des TI (Publié actuellement en anglais seulement)
109. ISO/IEC TR 13335-2:1997 Technologies de l'information -- Lignes directrices pour le management de sécurité IT -- Partie 2: Management et planning de sécurité IT (Publié actuellement en anglais seulement)
110. ISO/IEC DTR 13335-3 Titre manque
111. ISO/IEC 13522-1:1997 Technologies de l'information -- Codage de l'information multimédia et hypermédia -- Partie 1: Représentation d'objet MHEG -- Notation de base (ASN.1) (Publiée actuellement en anglais seulement)
112. ISO/IEC 13522-3:1997 Technologies de l'information -- Codage de l'information multimédia et hypermédia -- Partie 3: Représentation d'interéchange script MHEG (Publiée actuellement en anglais seulement)
113. ISO/IEC 13522-4:1996 Technologies de l'information -- Représentation codée d'informations multimédia et hypermédia -- Partie 4: Procédure d'enregistrement MHEG (Publiée actuellement en anglais seulement)

114. ISO/IEC 13522-5:1997 Technologies de l'information -- Codage de l'information multimédia et hypermédia -- Partie 5: Support pour applications interactives de niveau fondamental (Publiée actuellement en anglais seulement)
115. ISO/IEC DIS 13522-6 Technologies de l'information -- Codage de l'information multimédia et hypermédia -- Partie 6: Support pour les applications interactives améliorées
116. ISO/IEC 13818-1:1996 Technologies de l'information -- Codage générique des images animées et du son associé: Systèmes
117. ISO/IEC 13818-2:1996 Technologies de l'information -- Codage générique des images animées et du son associé: Données vidéo
118. ISO/IEC 13818-3:1995 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 3: Son (Publiée actuellement en anglais seulement)
119. ISO/IEC DIS 13818-3 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 3: Son (Révision de l'ISO/CEI 13818-3:1995) (DIS distribué en version anglaise seulement)
120. ISO/IEC DIS 13818-4 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 4: Essais de conformité
121. ISO/IEC TR 13818-5:1997 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 5: Simulation de logiciel (Publiée actuellement en anglais seulement)
122. ISO/IEC DIS 13818-6 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 6: Extensions pour DSM-CC (DIS distribué en version anglaise seulement)
123. ISO/IEC 13818-7:1997 Technologies de l'information -- Codage générique des images animées et du son associé -- Partie 7: Codage du son avancé (AAC) (Publiée actuellement en anglais seulement)
124. ISO/IEC 13818-9:1996 Technologies de l'information -- Codage générique des images animées et des informations sonores associées -- Partie 9: Extension pour interface temps réel pour systèmes décodeurs (Publiée actuellement en anglais seulement)
125. ISO/IEC 13888-1:1997 Technologies de l'information -- Techniques de sécurité -- Non-répudiation -- Partie 1: Généralités (Publiée actuellement en anglais seulement)
126. ISO/IEC DIS 13888-2 Technologies de l'information -- Techniques de sécurité -- Non-répudiation - - Partie 2: Mécanismes utilisant des techniques symétriques
127. ISO/IEC 13888-3:1997 Technologies de l'information -- Techniques de sécurité -- Non-répudiation -- Partie 3: Mécanismes utilisant des techniques asymétriques (Publiée actuellement en anglais seulement)
128. ISO/IEC DIS 14495-1 Titre manque
129. ISO/IEC DIS 14888-1 Traitement de l'information -- Techniques de sécurité -- Signatures digitales avec appendice -- Partie 1: Généralités
130. ISO/IEC DIS 14888-2 Titre manque

131. ISO/IEC DIS 14888-3 Titre manque
132. ISO/IEC 14957:1996 Technologies de l'information -- Notation du format pour les valeurs des éléments de données (Publiée actuellement en anglais seulement)
133. ISO/IEC DIS 14980 Technologies de l'information -- Code de pratique pour la gestion de la sécurité dans l'information
134. ISO/IEC 15200:1996 Technologies de l'information -- Algorithme de compression de données d'adaptation sans pertes (ALDC) (Publiée actuellement en anglais seulement)
135. ISO/DIS 15394 Emballage -- Codes à barres et symboles bidimensionnels pour l'expédition, le transport et les étiquettes de réception
136. ISO/IEC DIS 16022 Technologies de l'information -- Spécification internationale des symboles -- Matrice de données (DIS distribué en version anglaise seulement)
137. ISO/IEC DIS 16023 Technologies de l'information -- Spécification internationale des symboles -- MaxiCode (DIS distribué en version anglaise seulement)

ANNEXE B

Les codes de contrôle ASCII

Les codes de contrôle d'un codage 7 bits sont classés en 6 catégories. Une explication détaillée de chaque caractère de contrôle est mentionnée dans chaque norme ISO (Notre source : ISO 9036 : Information processing - Arabic 7-bit coded character set information interchange. Première édition, 1987-04-15, Référence iso 9036 : 1987 (E)).

Tab. 1 : Les caractères de contrôle pour la transmission

Acronyme	Nom de code	Position
SOH	Start of Heading	0/1
STX	Start of text	0/2
ETX	End of Text	0/3
EOT	End of transmission	0/4
ENQ	Enquiry	0/5
ACK	Acknowledge	0/6
DLE	Data Link Escape	1/0
NAK	Negative Acknowledge	1/5
SYN	Synchronous Idle	1/6
ETB	End of transmission Block	1/7

Tab. 2 : Les caractères de contrôle des paramètres de formatage

Acronyme	Nom de code	Position
BS	Backspace	0/8
HT	Horizontal Tabulation	0/9
LF	Line Feed	0/10
VT	Vertical Tabulation	0/11
FF	Form Feed	0/12
CR	Carriage return	0/13

Tab. 3 : Les caractères de contrôle des codes d'extension

Acronyme	Nom de code	Position
SO	Shift-Out	0/14
SI	Shift-In	0/15
ESC	Escape	1/11

Tab. 4 : Les caractères de contrôle de matériel

Acronyme	Nom de code	Position
DC1	Device Control 1	1/1
DC2	Device Control 2	1/2
DC3	Device Control 3	1/3
DC4	Device Control 4	1/4

Tab. 5 : Les caractères de contrôle des séparateurs d'information

Acronyme	Nom de code	Position
IS4	Information separator 4 (file separator)	1/12
IS3	Information separator 3 (group separator)	1/13
IS2	Information separator 2 (record separator)	1/14
IS1	Information separator 1 (unit separator)	1/15

Tab. 6 : Des caractères de contrôle divers

Acronyme	Nom de code	Position
NUL	Null	0/0
BEL	Bell	0/7
CAN	Cancel	1/8
EM	End of Medium	1/9
SUB	Substitute Character	1/10
DEL	Delete	7/15

ANNEXE C

Jeu de caractères ISO 646 : Représentation codée

N°	Glyphes	<escape_char> /	Nom du glyphe
1.	<NUL>	/d000	NULL (NUL)
2.	<SOH>	/d001	START OF HEADING (SOH)
3.	<STX>	/d002	START OF TEXT (STX)
4.	<ETX>	/d003	END OF TEXT (ETX)
5.	<EOT>	/d004	END OF TRANSMISSION (EOT)
6.	<ENQ>	/d005	ENQUIRY (ENQ)
7.	<ACK>	/d006	ACKNOWLEDGE (ACK)
8.	<alert>	/d007	BELL (BEL)
9.	<BEL>	/d007	BELL (BEL)
10.	<backspace>	/d008	BACKSPACE (BS)
11.	<tab>	/d009	CHARACTER TABULATION (HT)
12.	<newline>	/d010	LINE FEED (LF)
13.	<vertical-tab>	/d011	LINE TABULATION (VT)
14.	<form-feed>	/d012	FORM FEED (FF)
15.	<carriage-return>	/d013	CARRIAGE RETURN (CR)
16.	<DLE>	/d016 DATALINK ESCAPE (DLE)	
17.	<DC1>	/d017 DEVICE CONTROL ONE (DC1)	
18.	<DC2>	/d018 DEVICE CONTROL TWO (DC2)	
19.	<DC3>	/d019 DEVICE CONTROL THREE (DC3)	
20.	<DC4>	/d020 DEVICE CONTROL FOUR (DC4)	
21.	<NAK>	/d021 NEGATIVE ACKNOWLEDGE (NAK)	
22.	<SYN>	/d022 SYNCHRONOUS IDLE (SYN)	
23.	<ETB>	/d023 END OF TRANSMISSION BLOCK (ETB)	
24.	<CAN>	/d024 CANCEL (CAN)	
25.	<SUB>	/d026 SUBSTITUTE (SUB)	
26.	<ESC>	/d027 ESCAPE (ESC)	
27.	<IS4>	/d028 FILE SEPARATOR (IS4)	
28.	<IS3>	/d029 GROUP SEPARATOR (IS3)	
29.	<intro>	/d029 GROUP SEPARATOR	

		(IS3)	
30.	<IS2>	/d030 RECORD SEPARATOR (IS2)	
31.	<IS1>	/d031 UNIT SEPARATOR (IS1)	
32.	<space>	/d032	SPACE
33.	<exclamation-mark>	/d033	EXCLAMATION MARK
34.	<quotation-mark>	/d034	QUOTATION MARK
35.	<number-sign>	/d035	NUMBER SIGN
36.	<dollar-sign>	/d036	DOLLAR SIGN
37.	<percent-sign>	/d037	PERCENT SIGN
38.	<ampersand>	/d038	AMPERSAND
39.	<apostrophe>	/d039	APOSTROPHE
40.	<left-parenthesis>	/d040	LEFT PARENTHESIS
41.	<right-parenthesis>	/d041	RIGHT PARENTHESIS
42.	<asterisk>	/d042	ASTERISK
43.	<plus-sign>	/d043	PLUS SIGN
44.	<comma>	/d044	COMMA
45.	<hyphen>	/d045	HYPHEN-MINUS
46.	<hyphen-minus>	/d045	HYPHEN-MINUS
47.	<period>	/d046	FULL STOP
48.	<full-stop>	/d046	FULL STOP
49.	<slash>	/d047	SOLIDUS
50.	<solidus>	/d047	SOLIDUS
51.	<zero>	/d048	DIGIT ZERO
52.	<one>	/d049	DIGIT ONE
53.	<two>	/d050	DIGIT TWO
54.	<three>	/d051	DIGIT THREE
55.	<four>	/d052	DIGIT FOUR
56.	<five>	/d053	DIGIT FIVE
57.	<six>	/d054	DIGIT SIX
58.	<seven>	/d055	DIGIT SEVEN
59.	<eight>	/d056	DIGIT EIGHT
60.	<nine>	/d057	DIGIT NINE
61.	<colon>	/d058	COLON
62.	<semicolon>	/d059	SEMICOLON
63.	<less-than>	/d060	LESS-THAN SIGN
64.	<less-than-sign>	/d060	LESS-THAN SIGN
65.	<equals-sign>	/d061	EQUALS SIGN
66.	<equal-sign>	/d061	EQUALS SIGN
67.	<greater-than>	/d062	GREATER-THAN SIGN
68.	<greater-than-sign>	/d062	GREATER-THAN SIGN
69.	<question-mark>	/d063	QUESTION MARK
70.	<commercial-at>	/d064	COMMERCIAL AT
71.	<left-square-bracket>	/d091	LEFT SQUARE BRACKET
72.	<left-bracket>	/d091	LEFT SQUARE BRACKET
73.	<reverse-solidus>	/d092	REVERSE SOLIDUS

74.	<backslash>	/d092	REVERSE SOLIDUS
75.	<right-square-bracket>	/d093	RIGHT SQUARE BRACKET
76.	<right-bracket>	/d093	RIGHT SQUARE BRACKET
77.	<circumflex-accent>	/d094	CIRCUMFLEX ACCENT
78.	<low-line>	/d095	LOW LINE
79.	<underline>	/d095	LOW LINE
80.	<grave-accent>	/d096	GRAVE ACCENT
81.	<left-curly-bracket>	/d123	LEFT CURLY BRACKET
82.	<left-brace>	/d123	LEFT CURLY BRACKET
83.	<vertical-line>	/d124	VERTICAL LINE
84.	<right-curly-bracket>	/d125	RIGHT CURLY BRACKET
85.	<tilde>	/d126	TILDE
86.	<SP>	/d032	SPACE
87.	<!>	/d033	EXCLAMATION MARK
88.	<">	/d034	QUOTATION MARK
89.	<No>	/d035	NUMBER SIGN
90.	<Do>	/d036	DOLLAR SIGN
91.	<%>	/d037	PERCENT SIGN
92.	<&>	/d038	AMPERSAND
93.	<'>	/d039	APOSTROPHE
94.	<(>	/d040	LEFT PARENTHESIS
95.	<)>	/d041	RIGHT PARENTHESIS
96.	<*>	/d042	ASTERISK
97.	<+>	/d043	PLUS SIGN
98.	<, >	/d044	COMMA
99.	<- >	/d045	HYPHEN-MINUS
100.	<.>	/d046	FULL STOP
101.	<//>	/d047	SOLIDUS
102.	<0>	/d048	DIGIT ZERO
103.	<1>	/d049	DIGIT ONE
104.	<2>	/d050	DIGIT TWO
105.	<3>	/d051	DIGIT THREE
106.	<4>	/d052	DIGIT FOUR
107.	<5>	/d053	DIGIT FIVE
108.	<6>	/d054	DIGIT SIX
109.	<7>	/d055	DIGIT SEVEN
110.	<8>	/d056	DIGIT EIGHT
111.	<9>	/d057	DIGIT NINE
112.	<:>	/d058	COLON
113.	<;>	/d059	SEMICOLON
114.	<<>	/d060	LESS-THAN SIGN
115.	<=>	/d061	EQUALS SIGN
116.	<>>>	/d062	GREATER-THAN SIGN
117.	<?>	/d063	QUESTION MARK
118.	<At>	/d064	COMMERCIAL AT

119.	<A>	/d065	LATIN CAPITAL LETTER A
120.		/d066	LATIN CAPITAL LETTER B
121.	<C>	/d067	LATIN CAPITAL LETTER C
122.	<D>	/d068	LATIN CAPITAL LETTER D
123.	<E>	/d069	LATIN CAPITAL LETTER E
124.	<F>	/d070	LATIN CAPITAL LETTER F
125.	<G>	/d071	LATIN CAPITAL LETTER G
126.	<H>	/d072	LATIN CAPITAL LETTER H
127.	<I>	/d073	LATIN CAPITAL LETTER I
128.	<J>	/d074	LATIN CAPITAL LETTER J
129.	<K>	/d075	LATIN CAPITAL LETTER K
130.	<L>	/d076	LATIN CAPITAL LETTER L
131.	<M>	/d077	LATIN CAPITAL LETTER M
132.	<N>	/d078	LATIN CAPITAL LETTER N
133.	<O>	/d079	LATIN CAPITAL LETTER O
134.	<P>	/d080	LATIN CAPITAL LETTER P
135.	<Q>	/d081	LATIN CAPITAL LETTER Q
136.	<R>	/d082	LATIN CAPITAL LETTER R
137.	<S>	/d083	LATIN CAPITAL LETTER S
138.	<T>	/d084	LATIN CAPITAL LETTER T
139.	<U>	/d085	LATIN CAPITAL LETTER U
140.	<V>	/d086	LATIN CAPITAL LETTER V
141.	<W>	/d087	LATIN CAPITAL LETTER W
142.	<X>	/d088	LATIN CAPITAL LETTER X
143.	<Y>	/d089	LATIN CAPITAL LETTER Y
144.	<Z>	/d090	LATIN CAPITAL LETTER Z
145.	<<(>	/d091	LEFT SQUARE BRACKET
146.	<////>	/d092	REVERSE SOLIDUS
147.	<)>>>	/d093	RIGHT SQUARE BRACKET
148.	<'>>>	/d094	CIRCUMFLEX ACCENT
149.	<_>	/d095	LOW LINE
150.	<'!>	/d096	GRAVE ACCENT
151.	<a>	/d097	LATIN SMALL LETTER A
152.		/d098	LATIN SMALL LETTER B
153.	<c>	/d099	LATIN SMALL LETTER C
154.	<d>	/d100	LATIN SMALL LETTER D
155.	<e>	/d101	LATIN SMALL LETTER E
156.	<f>	/d102	LATIN SMALL LETTER F
157.	<g>	/d103	LATIN SMALL LETTER G
158.	<h>	/d104	LATIN SMALL LETTER H
159.	<i>	/d105	LATIN SMALL LETTER I
160.	<j>	/d106	LATIN SMALL LETTER J
161.	<k>	/d107	LATIN SMALL LETTER K
162.	<l>	/d108	LATIN SMALL LETTER L
163.	<m>	/d109	LATIN SMALL LETTER M
164.	<n>	/d110	LATIN SMALL LETTER N
165.	<o>	/d111	LATIN SMALL LETTER O
166.	<p>	/d112	LATIN SMALL LETTER P

167.	<q>	/d113	LATIN SMALL LETTER Q
168.	<r>	/d114	LATIN SMALL LETTER R
169.	<s>	/d115	LATIN SMALL LETTER S
170.	<t>	/d116	LATIN SMALL LETTER T
171.	<u>	/d117	LATIN SMALL LETTER U
172.	<v>	/d118	LATIN SMALL LETTER V
173.	<w>	/d119	LATIN SMALL LETTER W
174.	<x>	/d120	LATIN SMALL LETTER X
175.	<y>	/d121	LATIN SMALL LETTER Y
176.	<z>	/d122	LATIN SMALL LETTER Z
177.	<(!>	/d123	LEFT CURLY BRACKET
178.	<!!>	/d124	VERTICAL LINE
179.	<!)>	/d125	RIGHT CURLY BRACKET
180.	<'>	/d126	TILDE
181.	<NU>	/d000	NULL (NUL)
182.	<SH>	/d001	START OF HEADING (SOH)
183.	<SX>	/d002	START OF TEXT (STX)
184.	<EX>	/d003	END OF TEXT (ETX)
185.	<ET>	/d004	END OF TRANSMISSION (EOT)
186.	<EQ>	/d005	ENQUIRY (ENQ)
187.	<AK>	/d006	ACKNOWLEDGE (ACK)
188.	<BL>	/d007	BELL (BEL)
189.	<BS>	/d008	BACKSPACE (BS)
190.	<HT>	/d009	CHARACTER TABULATION (HT)
191.	<LF>	/d010	LINE FEED (LF)
192.	<VT>	/d011	LINE TABULATION (VT)
193.	<FF>	/d012	FORM FEED (FF)
194.	<CR>	/d013	CARRIAGE RETURN (CR)
195.	<SO>	/d014	SHIFT OUT (SO)
196.	<SI>	/d015	SHIFT IN (SI)
197.	<DL>	/d016	DATALINK ESCAPE (DLE)
198.	<D1>	/d017	DEVICE CONTROL ONE (DC1)
199.	<D2>	/d018	DEVICE CONTROL TWO (DC2)
200.	<D3>	/d019	DEVICE CONTROL THREE (DC3)
201.	<D4>	/d020	DEVICE CONTROL FOUR (DC4)
202.	<NK>	/d021	NEGATIVE ACKNOWLEDGE (NAK)
203.	<SY>	/d022	SYNCHRONOUS IDLE (SYN)
204.	<EB>	/d023	END OF TRANSMISSION BLOCK (ETB)
205.	<CN>	/d024	CANCEL (CAN)
206.		/d025	END OF MEDIUM (EM)
207.	<SB>	/d026	SUBSTITUTE (SUB)
208.	<EC>	/d027	ESCAPE (ESC)
209.	<FS>	/d028	FILE SEPARATOR (IS4)

210.	<GS>	/d029	GROUP SEPARATOR (IS3)
211.	<RS>	/d030	RECORD SEPARATOR (IS2)
212.	<US>	/d031	UNIT SEPARATOR (IS1)
213.	<DT>	/d127	DELETE (DEL)

ANNEXE D1

INFORMATIONS SUPPLEMENTAIRES SUR MS-DOS 6.22

PERSONNALISATION EN VUE D'UNE UTILISATION INTERNATIONALE

Ce fichier fournit des informations sur la personnalisation de MS-DOS 6.22 en vue d'une utilisation internationale. Ces informations ne figurent pas dans le GUIDE DE L'UTILISATEUR DE MICROSOFT MS-DOS 6.22, ni dans l'aide de MS-DOS ou le fichier LISEZMOI.TXT. Il complète les informations sur les jeux de caractères et les claviers internationaux inclus dans la section 1 du fichier LISEZMOI.TXT fourni avec MS-DOS 6.22.

Ce fichier aborde les sujets suivants :

1. Prise en charge de claviers et de jeux de caractères (pages de code) supplémentaires
 - 1.1 Fichier KEYBRD2.SYS
 - 1.2 Fichiers EGA2.CPI et EGA3.CPI
 - 1.3 Configuration de votre ordinateur pour l'utilisation des paramètres internationaux
2. Passage d'un mode à l'autre sur les claviers à deux modes
3. Utilisation des polices ISO
4. Tableaux récapitulatifs des langues prises en charge

1. Prise en charge de claviers et de jeux de caractères (pages de code) supplémentaires

MS-DOS 6.22 comprend les fichiers KEYBRD2.SYS, EGA2.CPI et EGA3.CPI.

Ces derniers permettent la prise en charge de claviers et de jeux de caractères (pages de code) supplémentaires. Vous pouvez utiliser ces fichiers de la même façon que les fichiers KEYBOARD.SYS et EGA.CPI.

MS-DOS comprend également des nouveaux paramètres de pays dans COUNTRY.SYS. (En outre, MS-DOS 6.22 inclut des versions compatibles ISO de certains jeux de caractères. Pour plus d'informations, reportez-vous à la section 3 de ce fichier.)

MS-DOS 6.22 comprend de nouveaux paramètres de pays dans COUNTRY.SYS qui permettent l'utilisation des nouveaux claviers et jeux de caractères. Pour plus d'informations, reportez-vous aux tableaux de la section 4 de ce fichier.

Pour obtenir la liste complète des pays, claviers et jeux de caractères fournis par MS-DOS 6.22, reportez-vous à la section 4 de ce fichier.

1.1 Fichier KEYBRD2.SYS

Le fichier KEYBRD2.SYS est très semblable au fichier KEYBOARD.SYS. Les principales différences entre les deux fichiers sont les suivantes :

o KEYBRD2.SYS offre un support pour les claviers suivants non gérés par KEYBOARD.SYS :

Pays/Langue	Code de clavier	ID clavier
Brésil (clavier standard secondaire)	br	274, 275
Bulgarie	bg	442
Canada (clavier standard CAN/CSA-Z)	cf	Voir ci-dessous
Grèce	gk	319
Islande	is	161
Macédoine et Serbie/Monténégro	yc	118
Roumanie	ro	333
Russie	ru	441
Turquie	tr	440, 179

o Bien que le fichier KEYBRD2.SYS contienne un clavier canadien doté du même code de clavier (058) et de la même ID (cf) que celui du fichier KEYBOARD.SYS, la disposition des touches sur les deux claviers est différente. Le clavier du fichier KEYBRD2.SYS répond aux normes définies par la spécification 243.200-92 (CAN/CSA-Z) du Gouvernement canadien, tandis que le clavier du fichier KEYBOARD.SYS, qui est le plus couramment utilisé, ne répond pas à ces normes.

o Le fichier KEYBRD2.SYS ne contient pas de support pour les claviers suivants, qui sont en revanche gérés par KEYBOARD.SYS :

Amérique latine
Belgique
Danemark
Espagne
France
Japon
Pays-Bas
Portugal
Suisse (Allemand)
Suisse (Français)

o KEYBRD2.SYS contient des modifications au clavier allemand qui ne figurent pas dans le fichier KEYBOARD.SYS. Sur le clavier allemand, la touche le verrouillage des majuscules située sur la rangée supérieure du clavier a la même fonction que sur le clavier américain (le mode "machine à écrire" a été supprimé).

1.2 Fichiers EGA2.CPI et EGA3.CPI

Les fichiers EGA2.CPI et EGA3.CPI sont semblables au fichier EGA.CPI, à part le fait que ces deux fichiers contiennent différentes pages de code.

EGA2.CPI contient les pages de code suivantes :

737 - Grec II
850 - Multilingue (Latin I)

852 - Pays slaves - Europe de l'Est (Latin II)
 857 - Turc
 861 - Islandais
 869 - Grec

EGA3.CPI contient les pages de code suivantes :

437 - Américain
 850 - Multilingue (Latin I)
 852 - Pays slaves - Europe de l'Est (Latin II)
 855 - Cyrillique I
 866 - Russe (Cyrillique II)

Vous utilisez les fichiers KEYBRD2.SYS, EGA2.CPI et EGA3.CPI de la même façon que les fichiers KEYBRD.SYS et EGA.CPI.

1.3 Configuration de votre ordinateur pour l'utilisation des paramètres internationaux

Pour utiliser les paramètres nationaux ou un jeu de caractères autre que celui par défaut (États-Unis), vous devez modifier vos fichiers CONFIG.SYS et AUTOEXEC.BAT.

Par exemple, pour utiliser les nouveaux paramètres islandais, procédez comme suit :

1. Ajoutez les commandes suivantes a votre fichier CONFIG.SYS :

```
REM PARAMETRES ISLANDE
COUNTRY=354,861,C:\DOS\COUNTRY.SYS
DEVICE=C:\DOS\DISPLAY.SYS CON=(,2)
```

2. Ajoutez les commandes suivantes a votre fichier AUTOEXEC.BAT :

```
REM PARAMETRES ISLANDE
MODE CON CP PREPARE=((861,850) C:\DOS\EGA2.CPI)
MODE CON CP SELECT=861
KEYB IS,,C:\DOS\KEYBOARD2.SYS
```

3. Redémarrez votre ordinateur.

2. Passage d'un mode a l'autre sur les claviers a deux modes

MS-DOS 6.22 prend en charge les claviers a deux modes pour les claviers grec, canadien français, bulgare, serbo-monténégain et russe. Avec ces claviers, vous pouvez passer aisément d'un mode a l'autre. Lorsque votre ordinateur démarre avec l'un de ces claviers, le mode actif est le mode principal. Pour passer au mode secondaire, utilisez les combinaisons de touches ci-dessous.

Combinaisons de touches pour le clavier grec :

Mode	Combinaison de touches

Mode principal Latin	ALT + MAJ GAUCHE
Mode secondaire Grec	ALT + MAJ DROITE

Combinaisons de touches pour le clavier canadien français :

Mode	Combinaison de touches
Mode principal Latin	CTRL + MAJ DROITE
Mode secondaire Latin II	CTRL + MAJ GAUCHE

Combinaisons de touches pour les claviers bulgare, russe, serbe et monténégrin :

Mode	Combinaison de touches
Mode principal Latin	ALT + MAJ GAUCHE
Mode secondaire Cyrillique	ALT + MAJ DROITE

3. Utilisation des polices ISO

Les polices ISO sont des polices d'affichage non proportionnelles spéciales qui, sur certains écrans, sont plus lisibles que les polices standard. (Ces polices ne fonctionnent qu'en mode 25 lignes ; elles ne prennent pas en charge les modes 43 et 50 lignes.)

MS-DOS 6.22 inclut le fichier ISO.CPI, qui contient les versions compatibles ISO des jeux de caractères (pages de code) suivants :

437	Américain
737	Grec II
850	Multilingue (Latin I)
852	Pays slaves - Europe de l'Est (Latin II)
855	Cyrillique I
857	Turc
860	Portugais
861	Islandais
863	Canadien-Français
865	Scandinave
866	Russe (Cyrillique II)
869	Grec I

Les polices ISO fournies avec MS-DOS 6.22 sont conformes à la Partie 3 de la norme ISO 9241. Cette norme, mise au point par l'ISO (Independent Standards Organization), décrit un ensemble de standards matériels et logiciels qui, selon des experts en ergonomie, peuvent réduire la fatigue oculaire et améliorer la lisibilité. Pour être en parfaite conformité avec cette norme, utilisez les polices ISO de MS-DOS avec un moniteur répondant à la norme. Si vous n'êtes pas sûr que votre écran soit conforme, contactez votre fournisseur.

3.1 Installation des polices ISO

Pour installer les polices ISO, vous devez ajouter ou modifier certaines commandes figurant dans vos fichiers AUTOEXEC.BAT et CONFIG.SYS. Si votre système est déjà configuré pour charger un autre jeu de caractères (page de code) que le jeu standard (américain), effectuez la procédure 1. Sinon, effectuez la procédure 2.

Procédure 1 -- Si votre système est déjà configuré pour charger un autre jeu de caractères que le jeu standard (américain) :

1. Dans votre fichier AUTOEXEC.BAT, modifiez la commande `MODE CON CODEPAGE PREPARE` de sorte qu'elle spécifie ISO.CPI à la place de votre fichier de page de code (.CPI) habituel.

Supposons, par exemple, que votre commande `MODE` actuelle charge le jeu de caractères portugais (860) et soit la suivante :

```
MODE CON CODEPAGE PREPARE=((860) C:\DOS\EGA.CPI)
```

Pour utiliser la version compatible ISO du jeu de caractères portugais, vous devez modifier cette commande pour qu'elle utilise le fichier ISO.CPI. Vous devez donc écrire la commande suivante :

```
MODE CON CODEPAGE PREPARE=((860) C:\DOS\ISO.CPI)
```

Procédure 2 -- Si vous utilisez le jeu de caractères américain standard :

1. Dans votre fichier CONFIG.SYS, ajoutez une commande `DEVICE` pour le fichier DISPLAY.SYS. Par exemple, si vous souhaitez utiliser le jeu de caractères matériel américain (437) et que le fichier DISPLAY.SYS se trouve dans le répertoire C:\DOS, ajoutez la commande suivante à votre fichier CONFIG.SYS :

```
DEVICE=C:\DOS\DISPLAY.SYS CON=(EGA,437,2)
```

Remarque : Le fichier DISPLAY.SYS utilise environ 12 ko de mémoire.

2. Dans votre fichier AUTOEXEC.BAT, ajoutez des commandes `MODE` pour charger et sélectionner les jeux de caractères (pages de code) appropriés. Par exemple, pour utiliser la version ISO du jeu de caractères américain (437), ajoutez les commandes suivantes :

```
MODE CON CODEPAGE PREPARE=((437) C:\DOS\ISO.CPI)
MODE CON CODEPAGE SELECT=437
```

4. Tableaux récapitulatifs des langues prises en charge

Le tableau suivant présente la liste de tous les pays pris en charge par MS-DOS 6.22, ainsi que les jeux de caractères, le fichier CPI et les codes de clavier correspondants.

Remarque : Pour utiliser les polices ISO, indiquez le fichier ISO.CPI au lieu du fichier. CPI indiqué pour le pays concerné. (Pour plus d'informations, reportez-vous à la section précédente.)

Pays	Code du pays	Jeux de caractères	Fichier CPI	Code de clavier	ID de clavier
(code de pays IBM)	035	855,850	EGA3.CPI	bg	442
Afrique du Sud	027	437,850	EGA.CPI	us	---
Albanie	355	852,850	EGA.CPI	--	---
Allemagne	049	850,437	EGA.CPI	gr	---
Amérique latine	003	850,437	EGA.CPI	la	---
Anglais international	061	437,850	EGA.CPI	--	---
Argentine	054	850,437	EGA.CPI	la	---
Australie	061	437,850	EGA.CPI	us	---
Autriche	043	850,437	EGA.CPI	gr	---
Belgique	032	850,437	EGA.CPI	be	---
Bosnie-Herzégovine	387	852,850	EGA.CPI	yu	234
Brésil	055	850,437	EGA.CPI	br	274, 275
Bulgarie	359	855,850	EGA3.CPI	bg	442
Bulgarie					
Canada (anglais)	004	850,863	EGA.CPI	cf	Voir section1.1
Canada (français)	002	850,863	EGA.CPI	cf	Voir section1.1
Chili	056	850,437	EGA.CPI	la	---
Colombie	057	850,437	EGA.CPI	la	---
Corée	082	437, 934*	EGA.CPI	--	---
Croatie	384	852,850	EGA.CPI	yu	234
Danemark	045	850,865	EGA.CPI	dk	---
Equateur	593	850,437	EGA.CPI	la	---
Espagne	034	850,437	EGA.CPI	sp	---
États-Unis	001	437,850	EGA.CPI	us	---
Finlande	358	850,437	EGA.CPI	su	---
France	033	850,437	EGA.CPI	fr	120, 189
Grèce	030, 850	869, 737,	EGA2.CPI	gk	319
Hong-Kong	852	437, **	EGA.CPI	--	---
Hongrie	036	852,850	EGA.CPI	hu	---
Inde	091	437, **	EGA.CPI	--	---
Irlande	353	850,437	EGA.CPI	uk	---
Islande	354	850,861	EGA2.CPI	is	161
Israël	972	850,862*	EGA.CPI	--	---
Italie	039	850,437	EGA.CPI	it	141, 142
Japon	081	437, 932*	EGA.CPI	jp	---
Macédoine	389	855,850	EGA3.CPI	yc	118
Malaisie	060	437, **	EGA.CPI	--	---
Mexique	052	850,437	EGA.CPI	la	---

Norvège	047	850,865	EGA.CPI	no	---
Nouvelle-Zélande	064	437,850	EGA.CPI	us	---
Pays-Bas	031	850,437	EGA.CPI	nl	---
Pologne	048	852,850	EGA.CPI	pl	---
Portugal	351	850,860	EGA.CPI	po	---
République populaire de Chine	086	437, 936*	EGA.CPI	--	---
République tchèque	042	852,850	EGA.CPI	cz	243
Roumanie	040	852,850	EGA.CPI	ro	333
Royaume-Uni	044	850,437	EGA.CPI	uk	166, 168
Russie	007, 852, 850, 437	866,855,	EGA3.CPI	ru	441
Serbie/Monténégro	381	855,850	EGA3.CPI	yc	118
Singapour	065	437, **	EGA.CPI	--	---
Slovaquie	421***	852,850	EGA.CPI	sl	245
Slovénie	386	852,850	EGA.CPI	yu	234
Suède	046	850,437	EGA.CPI	sv	---
Suisse	041	850,437	EGA.CPI	sf (français) sg (allemand)	
Taiwan	886	437, 938*	EGA.CPI	--	---
Turquie	090	857,850	EGA2.CPI	tr	440, 179
Vénézuela	058	850,437	EGA.CPI	la	---
Yougoslavie	038	852,850	EGA.CPI	yu	234
Yougoslavie (cyrillique)	038	855,850	EGA3.CPI	yc	118
Zone méridionale(Pays arabes)	785	850,864*	EGA.CPI	--	---

* Bien que le fichier COUNTRY.SYS prenne en charge les conventions nationales (date, heure, devise, etc.) correspondant aux pages de code 862 (hébreu), 864 (arabe), 932 (japonais), 934 (coréen), 936 (chinois de la République populaire de Chine) et 938 (taiwanais), ces jeux de caractères (pages de code) ne sont disponibles qu'avec les versions spécifiques de MS-DOS correspondantes.

** Hong-Kong, l'Inde, la Malaisie et Singapour ont été ajoutés au fichier COUNTRY.SYS pour permettre d'accéder aux paramètres de ces pays lorsque l'anglais international est utilisé. Si vous avez besoin des paramètres d'un de ces pays alors que vous utilisez la langue officielle correspondante, contactez la filiale Microsoft du pays concerné (Microsoft Hong-Kong Limited, Microsoft Systems Private Limited (Inde), Microsoft Malaysia Sdn Bhd ou Microsoft Singapore Pte Ltd).

*** Un nouveau code de pays, 421, a été attribué à la Slovaquie car, dans MS-DOS 6.22, les paramètres nationaux de la Slovaquie et de la République tchèque sont suffisamment différents

pour nécessiter des codes distincts. Ces deux pays ont encore le même indicatif téléphonique (042).

Le tableau suivant présente la liste des formats de date et d'heure associés aux codes des pays correspondants. Pour chaque pays, la colonne Format de date indique la façon dont MS-DOS affichera la date du 20 juillet 1994 et la colonne Format d'heure la façon dont MS-DOS affichera l'heure suivante : 21 heures 45 minutes et 25 secondes.

Pays	Code de pays	Format de date	Format d'heure
Afrique du Sud	027	94-07-20	21:45:25
Albanie	355	94-07-20	21:45:25
Allemagne	049	20.07.94	21:45:25
Amérique latine	003	20/07/94	9:45:25pm
Anglais international	061	20/07/94	9:45:25pm
Argentine	054	20/07/94	21.45.25
Australie	061	20/07/94	9:45:25
Autriche	043	20.07.94	21:45:25
Belgique	032	20/07/94	21:45:25
Bosnie-Herzégovine	387	94-07-20	21:45:25
Brésil	055	20/07/94	21:45:25
Bulgarie	359	20.07.94	21.45.25
Canada (anglais)	004	20/07/94	21:45:25
Canada (français)	002	94-07-20	21:45:25
Chili	056	20/07/94	21:45:25
Colombie	057	20/07/94	9:45:25pm
Corée	082	94-07-20	21:45:25
Croatie	384	94-07-20	21:45:25
Danemark	045	20-07-94	21:45:25
Espagne	034	20/07/94	21:45:25
Equateur	593	20/07/94	21:45:25
États-Unis	251	07/20/94	9:45:25pm
Finlande	358	20.07.94	21.45.25
France	033	20/07/94	21:45:25
Grèce	030	20/07/94	9:45:25pm
Hong-Kong	852	20/07/94	21:45:25
Hongrie	036	94.07.20	21.45.25
Inde	091	20/07/94	21:45:25
Irlande	353	20/07/94	21:45:25
Islande	354	94-07-20	21:45:25
Israël	972	20/07/94	21:45:25
Italie	039	20/07/94	21.45.25
Japon	081	94-07-20	9:45:25pm
Macédoine	389	94-07-20	21:45:25
Malaisie	060	20/07/94	9:45:25pm
Mexique	052	20/07/94	9:45:25pm
Norvège	047	20.07.94	21:45:25
Nouvelle-Zélande	064	20/07/94	21:45:25
Pays-Bas	031	20-07-94	21:45:25

Pologne	048	94.07.20	21:45:25
Portugal	351	20-07-94	21:45:25
République populaire de Chine	086	94.07.20	9:45:25pm
République tchèque	042	20.07.94	21.45.25
Roumanie	040	94-07-20	21:45:25
Royaume-Uni	044	20/07/94	21:45:25
Russie	007	20.07.94	21:45:25
Serbie/Monténégro	381	94-07-20	21:45:25
Singapour	065	20/07/94	9:45:25pm
Slovaquie	421	20.07.94	21.45.25
Slovénie	386	94-07-20	21:45:25
Suède	046	94-07-20	21.45.25
Suisse	041	20.07.94	21:45:25
Taiwan	886	94/07/20	9:45:25pm
Turquie	090	20/07/94	21:45:25
Vénézuela	058	20/07/94	9:45:25pm
Yugoslavie	038	94-07-20	21:45:25
Yugoslavie (cyrillique)	038	94-07-20	21:45:25
Zone méridionale(pays arabes)	785	20/07/94	9:45:25pm

ANNEXE D2

Les pages de codes et les spécificités de Microsoft Win32 NLS API

Le tableau suivant illustre une variété de pages de codes et les supports techniques qui leurs sont associés dans un environnement Windows 3.x. Certaines de ces pages de codes peuvent être utilisées au niveau de l'affichage ou des systèmes de fichiers (OEMCP) ou au niveau de l'interface utilisateur de Windows (ACP). D'autres sont utilisées uniquement pour des fonctions particulières comme la traduction de et vers les tables Unicode, les pages de codes Macintosh ou EBCDIC. Windows 95 par contre, dispose d'un support multilingue interne d'entrée et de sortie. Un seul ACP est actif à la fois dans une session Windows. Unicode n'est pas désigné dans ce tableau comme un élément ACP ni OEMCP. Il peut cependant être utilisé comme source de codage de caractères pour les interfaces utilisateur graphiques (GUI) et les applications Win32 bits sous Windows NT 3.1, Windows NT 3.5, et Win95.

Légende :

CP n°: Numéro de la page de code

Nom : Le nom canonique du jeu de caractères

ACP: Éligibilité pour ACP

OEMCP: Éligibilité pour OEMCP

CP n°	NOM	ACP	OEMCP	US/Eur. NT 3.1	Windows 95	NT 3.5
1200	Unicode (BMP de l'ISO 10646)			X		X
1250	Windows 3.1 Europe de l'Est	X		X	X	X
1251	Windows 3.1 Cyrillique	X		X	X	X
1252	Windows 3.1 US (ANSI)	X		X	X	X
1253	Windows 3.1 Grec	X		X	X	X
1254	Windows 3.1 Turc	X		X	X	X
1255	Hébreu	X			X	X
1256	Arabe	X			X	X
1257	Balte	X			X	X
437	MS-DOS États Unis		X	X	X	X
708	Arabe (ASMO 708)		X		X	X
709	Arabe (ASMO 449+,					
BCO N V4)		X		X	X	
710	Arabe (Arabe Transparent)		X		X	X
720	Arabe (Asmo Transparent)		X		X	X
737	Grec (ancien 437G)		X	X	X	X
755	Balte		X		X	X
850	MS-DOS Multilingue (Latin I)		X	X	X	X
852	MS-DOS Slave (Latin II)		X	X	X	X
855	IBM Cyrillique		X	X	X	X

857	IBM Turc		X	X	X	X
860	MS-DOS Portugais		X	X	X	X
861	MS-DOS Islandais		X	X	X	X
862	Hébreu		X		X	X
863	MS-DOS Canadien-Français		X	X	X	X
864	Arabe		X		X	X
865	MS-DOS Nordique		X	X	X	X
866	MS-DOS Russe (USSR)		X	X	X	X
869	IBM Grec Moderne		X	X	X	X
874	Thaï	X	X		X	X
932	Japonais	X	X	NT-J	X	NT-J
936	Chinois simplifié (PRC, Singapour)	X	X		X	X
949	Coréen	X	X		X	X
950	Chinois (Taiwan, Hong Kong)	X	X		X	X
1361	Coréen (Johab)		X			
1000 0	Macintosh Romain			X		X

Note : Ce document est traduit de l'original « Code page data in the Microsoft Win32 NLS API Specification » à l'adresse URL : <http://www.palnet.com/~jack/arabic/winapi.html>

ANNEXE E

Liste des normes ASMO et ISO ayant rapport avec la langue arabe

Normes ASMO

- **ASMO 445** (1982) : le code à 5 bits utilisé par les télé-imprimantes bilingues. Télex bilingue (arabe/latin),
- **ASMO 449** (1985) : Traitement de l'information. Jeu de caractères arabes codés à 7 éléments pour l'échange d'information,
- **ASMO 584** (1985) : Conversion entre le jeu de caractères arabes codés sur 7 bits (ASMO 449) et le code à 5 bits utilisé par les télé-imprimantes bilingues (ASMO 445),
- **ASMO 587** (1985) : transcodage de 5 à 7 moments,
- **ASMO 662** (1985) : Traitement de l'information. Jeu de caractères à 8 éléments pour l'échange d'informations,
- **ASMO 663** (1987?) : répartition des caractères sur le clavier pour terminaux arabes,
- **ASMO 708** (1986?) : Traitement de l'information - jeu de caractères arabe/latin codé à 8 éléments pour l'échange d'information,
- **ASMO 968** (1988) : Conversion entre le jeu de caractères arabes codés sur 8 bits (ASMO 662) et le code à 5 bits utilisés par les téléimprimantes bilingues (ASMO 445).

Normes ISO

- **ISO 233** : (1984) Documentation - Translittération des caractères arabes vers les caractères latins.
- **ISO 233-2** : (1993) Information and documentation - Translittération des caractères arabes vers les caractères latins. - Partie 2: Langue arabe : Translittération simplifiée.
- **ISO/DIS 11822** : Information and documentation - Extension du jeu de caractères codé de l'alphabet arabe pour l'échange de l'information bibliographique.
- **ISO 639** : (1988) Code pour la représentation des noms des langues. Selon ce standard, la langue arabe est codée "ar".

- **ISO 646** : la base de tous les standards internationaux pour le traitement et l'échange de l'information. Version Internationale de Références (IRV) de l'ISO 646.
- **ISO 6438** : (1983) Documentation - Jeu de caractères codés africains pour l'échange de l'information bibliographique.
- **ISO 2022** : Traitement de l'information multilingue par extension de codes.
- **ISO 9036** (=ASMO 449) (1987) : Traitement de l'information - jeux de caractères arabes codés à 7 éléments pour l'échange d'information.
- **ISO 8859-6** (=ASMO 449^E=ASMO 708) (1987) : Traitement de l'information - jeux de caractères graphiques codés sur un seul octet - Partie 6 : alphabet latin/arabe.
- **Unicode** : Traitement de l'information. Jeu de caractères universel codés sur pdeux octets.
- **ISO 10646** : Traitement de l'information. Jeu de caractères universel codés sur plusieurs octets. Elle sert de base à tous les standards internationaux pour le traitement et l'échange de l'information.

ANNEXE F

Jeu de caractères ASMO 449, Alias ISO 9036, Alias ISO-IR-89.

Représentation codée

N°	Glyphes	<escape_char> /	Nom du glyphe
1.	<NU>	/d000	NULL (NUL)
2.	<SH>	/d001	START OF HEA/DING (SOH)
3.	<SX>	/d002	START OF TEXT (STX)
4.	<EX>	/d003	EN/D OF TEXT (ETX)
5.	<ET>	/d004	EN/D OF TRANSMISSION (EOT)
6.	<EQ>	/d005	ENQUIRY (ENQ)
7.	<AK>	/d006	ACKNOWLEDGE (ACK)
8.	<BL>	/d007	BELL (BEL)
9.	<BS>	/d008	BACKSPACE (BS)
10.	<HT>	/d009	CHARACTER TABULATION (HT)
11.	<LF>	/d010	LINE FEE/D (LF)
12.	<VT>	/d011	LINE TABULATION (VT)
13.	<FF>	/d012	FORM FEE/D (FF)
14.	<CR>	/d013	CARRIAGE RETURN (CR)
15.	<SO>	/d014	SHIFT OUT (SO)
16.	<SI>	/d015	SHIFT IN (SI)
17.	</DL>	/d016	/DATALINK ESCAPE (/DLE)
18.	</D1>	/d017	/DEVICE CONTROL ONE (/DC1)
19.	</D2>	/d018	/DEVICE CONTROL TWO (/DC2)
20.	</D3>	/d019	/DEVICE CONTROL THREE (/DC3)
21.	</D4>	/d020	/DEVICE CONTROL FOUR (/DC4)
22.	<NK>	/d021	NEGATIVE ACKNOWLEDGE (NAK)
23.	<SY>	/d022	SYNCHRONOUS I/DLE (SYN)
24.	<EB>	/d023	EN/D OF TRANSMISSION BLOCK (ETB)
25.	<CN>	/d024	CANCEL (CAN)
26.		/d025	EN/D OF ME/DIUM (EM)
27.	<SB>	/d026	SUBSTITUTE (SUB)
28.	<EC>	/d027	ESCAPE (ESC)
29.	<FS>	/d028	FILE SEPARATOR (IS4)
30.	<GS>	/d029	GROUP SEPARATOR (IS3)
31.	<RS>	/d030	RECORD SEPARATOR (IS2)
32.	<US>	/d031	UNIT SEPARATOR (IS1)
33.	<SP>	/d032	SPACE
34.	<!>	/d033	EXCLAMATION MARK
35.	<">	/d034	QUOTATION MARK
36.	<Nb>	/d035	NUMBER SIGN
37.	<Cu>	/d036	CURRENCY SIGN
38.	<%>	/d037	PERCENT SIGN
39.	<&>	/d038	AMPERSAND
40.	<'>	/d039	APOSTROPHE

41.	<(>	/d040	LEFT PARENTHESIS
42.	<)>	/d041	RIGHT PARENTHESIS
43.	<*>	/d042	ASTERISK
44.	<+>	/d043	PLUS SIGN
45.	<, +>	/d044	ARABIC COMMA
46.	<->	/d045	HYPHEN-MINUS
47.	<.>	/d046	FULL STOP
48.	<///>	/d047	SOLI/DUS
49.	<0>	/d048	/DIGIT ZERO
50.	<1>	/d049	/DIGIT ONE
51.	<2>	/d050	/DIGIT TWO
52.	<3>	/d051	/DIGIT THREE
53.	<4>	/d052	/DIGIT FOUR
54.	<5>	/d053	/DIGIT FIVE
55.	<6>	/d054	/DIGIT SIX
56.	<7>	/d055	/DIGIT SEVEN
57.	<8>	/d056	/DIGIT EIGHT
58.	<9>	/d057	/DIGIT NINE
59.	<:>	/d058	COLON
60.	<; +>	/d059	ARABIC SEMICOLON
61.	<<>	/d060	LESS-THAN SIGN
62.	<=>	/d061	EQUALS SIGN
63.	</>>	/d062	GREATER-THAN SIGN
64.	<? +>	/d063	ARABIC QUESTION MARK
65.	<At>	/d064	COMMERCIAL AT
66.	<H'>	/d065	ARABIC LETTER HAMZA
67.	<aM>	/d066	ARABIC LETTER ALEF WITH MA/D/DA ABOVE
68.	<aH>	/d067	ARABIC LETTER ALEF WITH HAMZA ABOVE
69.	<wH>	/d068	ARABIC LETTER WAW WITH HAMZA ABOVE
70.	<ah>	/d069	ARABIC LETTER ALEF WITH HAMZA BELOW
71.	<yH>	/d070	ARABIC LETTER YEH WITH HAMZA ABOVE
72.	<a+>	/d071	ARABIC LETTER ALEF
73.	<b+>	/d072	ARABIC LETTER BEH
74.	<tm>	/d073	ARABIC LETTER TEH MARBUTA
75.	<t+>	/d074	ARABIC LETTER TEH
76.	<tk>	/d075	ARABIC LETTER THEH
77.	<g+>	/d076	ARABIC LETTER JEEM
78.	<hk>	/d077	ARABIC LETTER HAH
79.	<x+>	/d078	ARABIC LETTER KHAH
80.	</d+>	/d079	ARABIC LETTER /DAL
81.	</dk>	/d080	ARABIC LETTER THAL
82.	<r+>	/d081	ARABIC LETTER REH
83.	<z+>	/d082	ARABIC LETTER ZAIN

84.	<s+>	/d083	ARABIC LETTER SEEN
85.	<sn>	/d084	ARABIC LETTER SHEEN
86.	<c+>	/d085	ARABIC LETTER SA/D
87.	</d/d>	/d086	ARABIC LETTER /DA/D
88.	<tj>	/d087	ARABIC LETTER TAH
89.	<zH>	/d088	ARABIC LETTER ZAH
90.	<e+>	/d089	ARABIC LETTER AIN
91.	<i+>	/d090	ARABIC LETTER GHAIN
92.	<<(>	/d091	LEFT SQUARE BRACKET
93.	<////>	/d092	REVERSE SOLI/DUS
94.	<)/>>	/d093	RIGHT SQUARE BRACKET
95.	<'>>	/d094	CIRCUMFLEX ACCENT
96.	<_>	/d095	LOW LINE
97.	<++>	/d096	ARABIC TATWEEL
98.	<f+>	/d097	ARABIC LETTER FEH
99.	<q+>	/d098	ARABIC LETTER QAF
100.	<k+>	/d099	ARABIC LETTER KAF
101.	<l+>	/d100	ARABIC LETTER LAM
102.	<m+>	/d101	ARABIC LETTER MEEM
103.	<n+>	/d102	ARABIC LETTER NOON
104.	<h+>	/d103	ARABIC LETTER HEH
105.	<w+>	/d104	ARABIC LETTER WAW
106.	<j+>	/d105	ARABIC LETTER ALEF MAKSURA
107.	<y+>	/d106	ARABIC LETTER YEH
108.	<:+>	/d107	ARABIC FATHATAN
109.	<"+>	/d108	ARABIC /DAMMATAN
110.	<=+>	/d109	ARABIC KASRATAN
111.	<//+>	/d110	ARABIC FATHA
112.	<'>+	/d111	ARABIC /DAMMA
113.	<1+>	/d112	ARABIC KASRA
114.	<3+>	/d113	ARABIC SHA/D/DA
115.	<0+>	/d114	ARABIC SUKUN
116.	<(!>	/d123	LEFT CURLY BRACKET
117.	<!!>	/d124	VERTICAL LINE
118.	<!)>	/d125	RIGHT CURLY BRACKET
119.	<'>->	/d126	OVERLINE
120.	</DT>	/d127	/DELETE (/DEL)
121.	<NUL>	/d000	

ANNEXE G

ISO 10646-1

Répertoire des caractères BMP

Ce répertoire contient la liste des noms de tous les caractères nommés de la norme internationale ISO/CÉI 10646-1 (les mêmes que dans le standard Unicode). Les idéogrammes CJC (chinois-japonais-coréen) n'ont pas de nom, et ne sont donc pas repris ici.

- Rangée 00 : latin de base et supplément Latin-1
- Rangée 01 : latin étendu A et latin étendu B (partie)
- Rangée 02 : latin étendu B (fin), alphabet phonétique et modificateurs de lettres
- Rangée 03 : signes diacritiques jonctifs, grec, copte et symboles grecs
- Rangée 04 : cyrillique
- Rangée 05 : arménien et hébreu
- Rangée 06 : arabe
- Rangées 07 et 08 : inutilisées
- Rangée 09 : devanâgari et bengali
- Rangée 0A : pendjabi et goudjarati
- Rangée 0B : oriya et tamoul
- Rangée 0C : télougou et kannara
- Rangée 0D : malayalam
- Rangée 0E : thaï et lao
- Rangée 0F : inutilisée
- Rangée 10 : géorgien
- Rangée 11 : jamo hangul
- Rangées 12 à 1D : inutilisées
- Rangée 1E : latin étendu additionnel
- Rangée 1F : grec étendu
- Rangée 20 : ponctuation générale, exposants et indices, symboles monétaires et signes diacritiques jonctifs pour symboles
- Rangée 21 : symboles de type lettre, formes numériques et flèches
- Rangée 22 : opérateurs mathématiques
- Rangée 23 : signes techniques divers
- Rangée 24 : pictogrammes de commande, reconnaissance optique de caractères et alphanumériques encerclés
- Rangée 25 : filets, pavés et formes géométriques
- Rangée 26 : symboles divers
- Rangée 27 : casseau
- Rangées 28 à 2F : inutilisées

- Rangée 30 : symboles et ponctuation CJC, hiragana et katakana
- Rangée 31 : bopomofo, jamo de compatibilité hangul et divers CJC
- Rangée 32 : lettres et mois CJC encerclés
- Rangée 33 : compatibilité CJC
- Rangée 34 : hangul (partie)
- Rangée 35 : hangul (partie)
- Rangée 36 : hangul (partie)
- Rangée 37 : hangul (partie)
- Rangée 38 : hangul (partie)
- Rangée 39 : hangul (partie)
- Rangée 3A : hangul (partie)
- Rangée 3B : hangul (partie)
- Rangée 3C : hangul (partie)
- Rangée 3D : hangul (fin) et hangul supplémentaire A (partie)
- Rangée 3E : hangul supplémentaire A (partie)
- Rangée 3F : hangul supplémentaire A (partie)
- Rangée 40 : hangul supplémentaire A (partie)
- Rangée 41 : hangul supplémentaire A (partie)
- Rangée 42 : hangul supplémentaire A (partie)
- Rangée 43 : hangul supplémentaire A (partie)
- Rangée 44 : hangul supplémentaire A (fin) et hangul supplémentaire B (partie)
- Rangée 45 : hangul supplémentaire B (partie)
- Rangée 46 : hangul supplémentaire B (partie)
- Rangée 47 : hangul supplémentaire B (partie)
- Rangée 48 : hangul supplémentaire B (partie)
- Rangée 49 : hangul supplémentaire B (partie)
- Rangée 4A : hangul supplémentaire B (partie)
- Rangée 4B : hangul supplémentaire B (partie)
- Rangée 4C : hangul supplémentaire B (partie)
- Rangée 4D : hangul supplémentaire B (fin)
- Rangées 4E à 9F : idéogrammes unifiés CJC
- Rangées A0 à D7 : inutilisées
- Rangées D8 à DF : réservées pour UTF-16
- Rangées E0 à F8 : zone à usage privé
- Rangées F9 à FA : formes de compatibilité CJC
- Rangée FB : formes de présentation alphabétiques et formes de présentation arabes-A (partie)
- Rangée FC : formes de présentation arabes-A (partie)
- Rangée FD : formes de présentation arabes-A (fin)
- Rangée FE : demi-signes jonctifs, formes de compatibilité CJC, petites variantes de forme et formes de présentation arabes-B
- Rangée FF : formes de demi et pleine chasse et caractères spéciaux

ANNEXE H

La table de correspondance entre ISO 8859-6 (caractères arabes) et UNICODE

#ISO 8859-6	#UNICODE	Nom du caractère
0x20	0x0020	SPACE
0x21	0x0021	EXCLAMATION MARK
0x22	0x0022	QUOTATION MARK
0x23	0x0023	NUMBER SIGN
0x24	0x0024	DOLLAR SIGN
0x25	0x0025	PERCENT SIGN
0x26	0x0026	AMPERSAND
0x27	0x0027	APOSTROPHE
0x28	0x0028	LEFT PARENTHESIS
0x29	0x0029	RIGHT PARENTHESIS
0x2A	0x002A	ASTERISK
0x2B	0x002B	PLUS SIGN
0x2C	0x002C	COMMA
0x2D	0x002D	HYPHEN-MINUS
0x2E	0x002E	FULL STOP
0x2F	0x002F	SOLIDUS
0x30	0x0660	ARABIC-INDIC DIGIT ZERO
0x31	0x0661	ARABIC-INDIC DIGIT ONE
0x32	0x0662	ARABIC-INDIC DIGIT TWO
0x33	0x0663	ARABIC-INDIC DIGIT THREE
0x34	0x0664	ARABIC-INDIC DIGIT FOUR
0x35	0x0665	ARABIC-INDIC DIGIT FIVE
0x36	0x0666	ARABIC-INDIC DIGIT SIX
0x37	0x0667	ARABIC-INDIC DIGIT SEVEN
0x38	0x0668	ARABIC-INDIC DIGIT EIGHT
0x39	0x0669	ARABIC-INDIC DIGIT NINE
0x3A	0x003A	COLON
0x3B	0x003B	SEMICOLON
0x3C	0x003C	LESS-THAN SIGN
0x3D	0x003D	EQUALS SIGN
0x3E	0x003E	GREATER-THAN SIGN
0x3F	0x003F	QUESTION MARK
0x40	0x0040	COMMERCIAL AT
0x41	0x0041	LATIN CAPITAL LETTER A
0x42	0x0042	LATIN CAPITAL LETTER B
0x43	0x0043	LATIN CAPITAL LETTER C
0x44	0x0044	LATIN CAPITAL LETTER D
0x45	0x0045	LATIN CAPITAL LETTER E
0x46	0x0046	LATIN CAPITAL LETTER F

0x47	0x0047	LATIN CAPITAL LETTER G
0x48	0x0048	LATIN CAPITAL LETTER H
0x49	0x0049	LATIN CAPITAL LETTER I
0x4A	0x004A	LATIN CAPITAL LETTER J
0x4B	0x004B	LATIN CAPITAL LETTER K
0x4C	0x004C	LATIN CAPITAL LETTER L
0x4D	0x004D	LATIN CAPITAL LETTER M
0x4E	0x004E	LATIN CAPITAL LETTER N
0x4F	0x004F	LATIN CAPITAL LETTER O
0x50	0x0050	LATIN CAPITAL LETTER P
0x51	0x0051	LATIN CAPITAL LETTER Q
0x52	0x0052	LATIN CAPITAL LETTER R
0x53	0x0053	LATIN CAPITAL LETTER S
0x54	0x0054	LATIN CAPITAL LETTER T
0x55	0x0055	LATIN CAPITAL LETTER U
0x56	0x0056	LATIN CAPITAL LETTER V
0x57	0x0057	LATIN CAPITAL LETTER W
0x58	0x0058	LATIN CAPITAL LETTER X
0x59	0x0059	LATIN CAPITAL LETTER Y
0x5A	0x005A	LATIN CAPITAL LETTER Z
0x5B	0x005B	LEFT SQUARE BRACKET
0x5C	0x005C	REVERSE SOLIDUS
0x5D	0x005D	RIGHT SQUARE BRACKET
0x5E	0x005E	CIRCUMFLEX ACCENT
0x5F	0x005F	LOW LINE
0x60	0x0060	GRAVE ACCENT
0x61	0x0061	LATIN SMALL LETTER A
0x62	0x0062	LATIN SMALL LETTER B
0x63	0x0063	LATIN SMALL LETTER C
0x64	0x0064	LATIN SMALL LETTER D
0x65	0x0065	LATIN SMALL LETTER E
0x66	0x0066	LATIN SMALL LETTER F
0x67	0x0067	LATIN SMALL LETTER G
0x68	0x0068	LATIN SMALL LETTER H
0x69	0x0069	LATIN SMALL LETTER I
0x6A	0x006A	LATIN SMALL LETTER J
0x6B	0x006B	LATIN SMALL LETTER K
0x6C	0x006C	LATIN SMALL LETTER L
0x6D	0x006D	LATIN SMALL LETTER M
0x6E	0x006E	LATIN SMALL LETTER N
0x6F	0x006F	LATIN SMALL LETTER O
0x70	0x0070	LATIN SMALL LETTER P
0x71	0x0071	LATIN SMALL LETTER Q
0x72	0x0072	LATIN SMALL LETTER R
0x73	0x0073	LATIN SMALL LETTER S
0x74	0x0074	LATIN SMALL LETTER T
0x75	0x0075	LATIN SMALL LETTER U
0x76	0x0076	LATIN SMALL LETTER V

0x77	0x0077	LATIN SMALL LETTER W
0x78	0x0078	LATIN SMALL LETTER X
0x79	0x0079	LATIN SMALL LETTER Y
0x7A	0x007A	LATIN SMALL LETTER Z
0x7B	0x007B	LEFT CURLY BRACKET
0x7C	0x007C	VERTICAL LINE
0x7D	0x007D	RIGHT CURLY BRACKET
0x7E	0x007E	TILDE
0xA0	0x00A0	NO-BREAK SPACE
0xA4	0x00A4	CURRENCY SIGN
0xAC	0x060C	ARABIC COMMA
0xAD	0x00AD	SOFT HYPHEN
0xBB	0x061B	ARABIC SEMICOLON
0xBF	0x061F	ARABIC QUESTION MARK
0xC1	0x0621	ARABIC LETTER HAMZA
0xC2	0x0622	ARABIC LETTER ALEF WITH MADDA ABOVE
0xC3	0x0623	ARABIC LETTER ALEF WITH HAMZA ABOVE
0xC4	0x0624	ARABIC LETTER WAW WITH HAMZA ABOVE
0xC5	0x0625	ARABIC LETTER ALEF WITH HAMZA BELOW
0xC6	0x0626	ARABIC LETTER YEH WITH HAMZA ABOVE
0xC7	0x0627	ARABIC LETTER ALEF
0xC8	0x0628	ARABIC LETTER BEH
0xC9	0x0629	ARABIC LETTER TEH MARBUTA
0xCA	0x062A	ARABIC LETTER TEH
0xCB	0x062B	ARABIC LETTER THEH
0xCC	0x062C	ARABIC LETTER JEEM
0xCD	0x062D	ARABIC LETTER HAH
0xCE	0x062E	ARABIC LETTER KHAH
0xCF	0x062F	ARABIC LETTER DAL
0xD0	0x0630	ARABIC LETTER THAL
0xD1	0x0631	ARABIC LETTER REH
0xD2	0x0632	ARABIC LETTER ZAIN
0xD3	0x0633	ARABIC LETTER SEEN
0xD4	0x0634	ARABIC LETTER SHEEN
0xD5	0x0635	ARABIC LETTER SAD
0xD6	0x0636	ARABIC LETTER DAD
0xD7	0x0637	ARABIC LETTER TAH
0xD8	0x0638	ARABIC LETTER ZAH
0xD9	0x0639	ARABIC LETTER AIN
0xDA	0x063A	ARABIC LETTER GHAIN
0xE0	0x0640	ARABIC TATWEEL
0xE1	0x0641	ARABIC LETTER FEH
0xE2	0x0642	ARABIC LETTER QAF

0xE3	0x0643	ARABIC LETTER KAF
0xE4	0x0644	ARABIC LETTER LAM
0xE5	0x0645	ARABIC LETTER MEEM
0xE6	0x0646	ARABIC LETTER NOON
0xE7	0x0647	ARABIC LETTER HEH
0xE8	0x0648	ARABIC LETTER WAW
0xE9	0x0649	ARABIC LETTER ALEF MAKSURA
0xEA	0x064A	ARABIC LETTER YEH
0xEB	0x064B	ARABIC FATHATAN
0xEC	0x064C	ARABIC DAMMATAN
0xED	0x064D	ARABIC KASRATAN
0xEE	0x064E	ARABIC FATHA
0xEF	0x064F	ARABIC DAMMA
0xF0	0x0650	ARABIC KASRA
0xF1	0x0651	ARABIC SHADDA
0xF2	0x0652	ARABIC SUKUN

ANNEXE I

Les rangées FB, FC, FD, FE de la BMP d'UNICODE réservée aux formes de présentation des caractères arabes surimprimés

	FB5	FB6	FB7	FB8	FB9	FBA	FBB	FBC		FBD	FBE	FBF	FC0	FC1	FC2	FC3	FC4
0	أ FB50	إ FB60	ق FB70	چ FB80	ك FB90	ن FBA0	ه FBB0		0		و FBE0	و FBF0	ج FC00	تي FC10	صح FC20	فم FC30	لح FC40
1	أ FB51	إ FB61	ق FB71	چ FB81	ك FB91	ن FBA1	ه FBB1		1		و FBE1	و FBF1	ج FC01	تي FC11	صح FC21	فم FC31	لح FC41
2	پ FB52	ٹ FB62	ج FB72	د FB82	گ FB92	ا FBA2			2		و FBE2	و FBF2	م FC02	ثم FC12	صح FC22	ني FC32	لم FC42
3	ب FB53	ت FB63	ج FB73	د FB83	گ FB93	ا FBA3			3	ث FBD3	و FBE3	و FBF3	ني FC03	ثي FC13	صح FC23	ني FC33	لي FC43
4	ب FB54	ت FB64	ج FB74	د FB84	گ FB94	ا FBA4			4	لث FBD4	ي FBE4	و FBF4	ني FC04	ثي FC14	صح FC24	فم FC34	لي FC44
5	ب FB55	ت FB65	ج FB75	د FB85	گ FB95	ا FBA5			5	س FBD5	ي FBE5	و FBF5	ج FC05	صح FC15	صح FC25	ني FC35	ج FC45
6	پ FB56	ٹ FB66	ج FB76	د FB86	گ FB96	ا FBA6			6	ك FBD6	ب FBE6	ي FBF6	ج FC06	ثم FC16	صح FC26	ني FC36	ج FC46
7	پ FB57	ٹ FB67	ج FB77	د FB87	گ FB97	ا FBA7			7	و FBD7	ب FBE7	ي FBF7	ج FC07	صح FC17	صح FC27	كا FC37	ج FC47
8	ب FB58	ت FB68	ج FB78	د FB88	گ FB98	ا FBA8			8	و FBD8	د FBE8	ب FBF8	م FC08	ثم FC18	صح FC28	كج FC38	م FC48
9	ب FB59	ت FB69	ج FB79	د FB89	گ FB99	ا FBA9			9	و FBD9	ب FBE9	ي FBF9	ني FC09	صح FC19	صح FC29	كج FC39	ي FC49
A	پ FB5A	ٹ FB6A	ج FB7A	د FB8A	گ FB9A	ا FBAA			A	و FBDa	نا FBEa	ي FBFa	بي FC0A	صح FC1A	صح FC2A	كج FC3A	ي FC4A
B	پ FB5B	ٹ FB6B	ج FB7B	د FB8B	گ FB9B	ا FBAB			B	و FBDB	نا FBEb	ب FBFb	ج FC0B	ثم FC1B	صح FC2B	كج FC3B	ج FC4B
C	ب FB5C	ت FB6C	ج FB7C	د FB8C	گ FB9C	ا FBAC			C	و FBDc	ب FBEc	ي FBFc	ج FC0C	صح FC1C	صح FC2C	ك FC3C	ج FC4C
D	ب FB5D	ت FB6D	ج FB7D	د FB8D	گ FB9D	ا FBAD			D	و FBDd	ب FBEd	ي FBFd	ج FC0D	صح FC1D	صح FC2D	كي FC3D	ج FC4D
E	ب FB5E	ت FB6E	ج FB7E	د FB8E	ن FB9E	ا FBAE			E	و FBDE	و FBEE	ب FBFE	ثم FC0E	صح FC1E	صح FC2E	كي FC3E	ثم FC4E
F	ب FB5F	ت FB6F	ج FB7F	د FB8F	ن FB9F	ا FBAF			F	و FBDF	و FBEF	ب FBFF	تي FC0F	ثم FC1F	صح FC2F	لح FC3F	ني FC4F

	FC5	FC6	FC7	FC8	FC9	FCA	FCB	FCC		FCD	FCE	FCF	FD0	FD1	FD2	FD3	FD4
0	ني FC50	ة FC60	تر FC70	كا FC80	نى FC90	ه FCA0	ص FCB0	خ FCC0	0	خ FCD0	م FCE0	ر FCF0	حي FD00	ضر FD10	بني FD20	شم FD30	
1	ش FC51	ه FC61	تر FC71	كي FC81	ير FC91	تج FCA1	صه FCB1	ف FCC1	1	م FCD1	ر FCE1	م FCF1	جى FD01	طى FD11	صى FD21	سم FD31	
2	هم FC52	ه FC62	تم FC72	كم FC82	يز FC92	تخ FCA2	صه FCB2	ن FCC2	2	ن FCD2	م FCE2	ا FCF2	جى FD02	طى FD12	صى FD22	شم FD32	
3	هى FC53	ه FC63	تى FC73	كى FC83	يم FC93	تخ FCA3	صه FCB3	ق FCC3	3	خ FCD3	د FCE3	ا FCF3	نخى FD03	هى FD13	ضى FD23	طه FD33	
4	هى FC54	نر FC64	تى FC74	كي FC84	ينى FC94	نم FCA4	صه FCB4	ك FCC4	4	خ FCD4	م FCE4	ا FCF4	نخى FD04	هى FD14	ضى FD24	سبه FD34	
5	مخ FC55	نر FC65	تي FC75	لم FC85	بى FC95	نم FCA5	صه FCB5	ك FCC5	5	ن FCD5	د FCE5	طى FCF5	صى FD05	نخى FD15	شبع FD25	سبه FD35	
6	مخ FC56	م FC66	تر FC76	لى FC86	يى FC96	نم FCA6	صه FCB6	ك FCC6	6	ن FCD6	م FCE6	طى FCF6	صى FD06	نخى FD16	شبع FD26	سبه FD36	
7	مخ FC57	تى FC67	تر FC77	لي FC87	بج FC97	نم FCA7	صه FCB7	ك FCC7	7	م FCD7	س FCE7	ع FCF7	صى FD07	سبى FD17	شبع FD27	شبع FD37	
8	م FC58	تى FC68	ثم FC78	ما FC88	نخ FC98	نم FCA8	صه FCB8	ك FCC8	8	م FCD8	س FCE8	ع FCF8	صى FD08	سبى FD18	شبع FD28	شبع FD38	
9	نخى FC59	تي FC69	تى FC79	مس FC89	نخ FC99	نم FCA9	صه FCB9	ل FCC9	9	ن FCD9	م FCE9	نخى FCF9	شبع FD09	شبع FD19	شبع FD29	شبع FD39	
A	نخى FC5A	بر FC6A	تى FC7A	نر FC8A	نم FC9A	نم FCAA	صه FCBA	ل FCCA	A	ن FCDa	م FCEa	نخى FCFa	شبع FD0A	شبع FD1A	س FD2A	طه FD3A	
B	ن FC5B	بز FC6B	تي FC7B	نر FC8B	نم FC9B	نم FCAb	صه FCBb	ل FCCb	B	ن FCDb	م FCEb	ك FCFb	شبع FD0b	سبى FD1b	ضر FD2b	طه FD3b	
C	ر FC5C	م FC6C	نى FC7C	نم FC8C	بج FC9C	نم FCAc	صه FCBc	ل FCCc	C	ن FCDc	م FCEc	سبى FCFc	شبع FD0c	سبى FD1c	ضر FD2c	طه FD3c	
D	نخى FC5D	نخى FC6D	نخى FC7D	نخى FC8D	بج FC9D	نم FCAd	صه FCBd	ل FCCd	D	ن FCDd	م FCEd	شبع FCFd	شبع FD0d	سبى FD1d	شبع FD2d	طه FD3d	
E	ن FC5E	بى FC6E	تى FC7E	نى FC8E	بج FC9E	نم FCAE	صه FCBE	ل FCCe	E	ن FCDe	م FCEe	شبع FCFe	س FD0E	سبى FD1E	شبع FD2E	طه FD3E	
F	ن FC5F	بى FC6F	تي FC7F	نى FC8F	ب FC9F	نم FCAf	صه FCBf	ل FCCf	F	ن FCDf	م FCEf	م FCFf	ضر FD0f	نخى FD1f	شبع FD2f	طه FD3f	

FD5		FD6		FD7		FD8		FD9		FDA		FDB		FDC		FDD		FDE		FDF	
0	نہج	مہ	نہج	لہم		نہج	مہ	نہج		نہج	مہ	نہج								صلی	
1	نہج	مہ	نہج	لہج		نہج	مہ	نہج		نہج	مہ	نہج								FDF0	
2	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								قلی	
3	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF1	
4	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								اللہ	
5	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF2	
6	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								اکبر	
7	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF3	
8	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								محمد	
9	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF4	
A	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								صلعم	
B	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF5	
C	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								رسول	
D	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF6	
E	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								علیہ	
F	نہج	مہ	نہج	لہج	نہج	نہج	مہ	نہج		نہج	مہ	نہج								FDF7	
																				وسلم	
																				FDF8	
																				صلی	
																				FDF9	
																				صلی اللہ علیہ وسلم	
																				FDFA	
																				صلی اللہ علیہ وسلم	
																				FDFB	
														</							