

RAZEQ AFZALI

ALEXANDRE ANDREEWSKI

Traitement automatique de la vocalisation en dari

Mathématiques et sciences humaines, tome 115 (1991), p. 67-75

http://www.numdam.org/item?id=MSH_1991__115__67_0

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1991, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (<http://msh.revues.org/>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

TRAITEMENT AUTOMATIQUE DE LA VOCALISATION EN DARI ¹

Razeq AFZALI ²
Alexandre ANDREEWSKI ³

RÉSUMÉ — *On étudie dans ce travail, grâce à un traitement automatique, le problème de la reconstitution vocalique à partir de l'écriture traditionnelle du dari (persan d'Afghanistan).*

On montre en particulier que cette écriture est très discriminante et qu'à partir d'une forme consonantique donnée, la forme vocalique correspondante est fortement déterminée.

On décrit la méthodologie générale suivie et on donne les statistiques correspondantes pour un lexique de taille moyenne et en voie de développement.

Les résultats de notre démarche corroborent notre hypothèse de travail : en effet, la différence entre oral et écrit est telle que, quand l'on passe de l'oral à l'écrit, la performance individuelle joue un rôle pertinent face à un texte donné ambigu graphiquement en persan.

SUMMARY — Automatic data processing of the vocalization in dari (persian language of Afghanistan).

In this article, is studied, thanks to an automatic data processing, the problem of the reconstitution of the vowel system, from the traditional writing of the dari (persian language of Afghanistan).

It's shown that this writing is very discriminant and that from a given consonantal form the corresponding vocalic form is strongly definite.

The methodology is described and we give the corresponding statistics for a medium size and developing lexicon. The results of our study corroborate our hypothesis : the individual performance, when we pass from oral to writing, has a pertinent part facing a given text graphically ambiguous.

1. INTRODUCTION

Le dari (persan d'Afghanistan) est une des langues indo-européennes accusatives du type indo-aryen dont l'écriture (voir III) n'explicite que les consonnes et les semi-voyelles et les voyelles longues mais pas les voyelles brèves. Dans tout ce qui suit, nous désignerons cette écriture par le sigle [C,SV,VL] (où C désigne les consonnes, SV la seule semi-voyelle et VL les voyelles longues) (voir III). Bien que les voyelles brèves ne soient pas écrites, pour la commodité de

¹ (Litt. "langue de la cour"), c'est une appellation pour distinguer le persan en usage en Afghanistan, du persan d'Iran. De ce fait, c'est la langue de la littérature, de l'Université, des mass medias et de l'administration. A ce niveau, elle ne diffère du persan d'Iran que par des faits lexicaux et, dans une moindre mesure, morphosyntaxiques, qui ne gênent pas l'intercompréhension.

² Docteur en linguistique automatique, chargé de cours à Paris IV et à l'UTC, a travaillé, en tant que chercheur associé, sur les problèmes de la traduction automatique au CELTA-CNRS (Nancy) et poursuit actuellement ses recherches en linguistique générale orientée vers les sciences cognitives et la linguistique computationnelle au CAMS-EHESS-Paris-Sorbonne.

³ Docteur es-sciences, travaille sur les problèmes de linguistique informatique. Il est père (avec C. FLUHR) du système d'indexation automatique et d'Interrogation en langage naturel. Il travaille actuellement au L.I.M.S.I. (CNRS-Orsay)

l'exposé, nous désignerons par [C,SV,VL,VB] (VB désigne les voyelles brèves), une écriture qui expliciterait les voyelles brèves. Nous soulignons que la notation [C,SV,VL,VB] est purement une facilité théorique.

Le dari s'écrit de droite à gauche. Pour sa transcription, on a adjoint aux 28 graphèmes arabes, 4 signes supplémentaires représentant des phonèmes proprement persans : /P/, /&/, /#/ , /G/. Les inscriptions du vieux perse, ancêtre du dari, sont gravées à l'aide d'une écriture cunéiforme créée uniquement à cet usage et qui s'écrit de gauche à droite. Syllabo-alphabétique, elle a été dérivée du syllabaire akkadien et comporte trente-sept signes, dont un pour marquer la séparation, auxquels s'ajoutent quelques idéogrammes ⁴.

La reconstitution des voyelles brèves à partir d'un graphisme [C,SV,VL] c'est à dire le passage de [C,SV,VL] à [C,SV,VL,VB] exige de la part du lecteur un effort lexical, syntaxique et sémantique qu'il réalise en tenant compte du contexte et de sa culture générale. C'est dans le but de pouvoir évaluer, dans cet effort, la part respective du lexique, de la syntaxe, de la sémantique, que nous avons entrepris cette étude.

2. HYPOTHÈSE DE TRAVAIL

Dans ce qui suit, seul l'aspect morphologique de la reconstitution vocalique est étudié et ce, non pas à partir du graphisme complet [C,SV,VL] qui est explicitement écrit et qui est, très souvent, non ambigu, mais à partir d'un graphisme de consonne et de semi-voyelle ordonnées [C,SV]. Ici, il y a lieu de souligner que C désigne la consonne mais SV désigne la semi-voyelle qui, comme on sait, intervient parfois comme voyelle et parfois comme consonne. Mais dans notre notation, le groupe [C,SV], sera toujours considéré comme constitué de consonnes. Le graphisme [C,SV] est bien entendu plus simple, quant à son écriture, mais beaucoup plus complexe quant à sa reconstitution vocalique. Dans notre travail, ce graphisme [C,SV] a un caractère purement théorique, puisque les voyelles longues sont toujours explicitement écrites et en tant qu'outil théorique, il permet d'approfondir la relation entre le lexique et sa représentation graphique. L'étude de ce graphisme n'est cependant pas fortuite, car, en persan, l'expérience des spécialistes montre qu'une écriture du type [C,SV] permet aussi, dans une large mesure, une identification du texte. Par conséquent, la réduction du graphisme [C,SV,VL] au graphisme [C,SV] bien qu'artificielle, permet d'effectuer d'une façon plus approfondie la relation entre lexique, système consonantique et système vocalique.

Illustrons, à l'aide d'un certain nombre d'exemples, notre hypothèse de travail.

1.1. Correspondance biunivoque entre lexique et graphisme [C,SV]

Il existe des mots comme /L^δKR/ graphisme de la forme (C.C.C.C.) du type [C,SV] qui correspondent d'une manière biunivoque au même paradigme lexical, déterminé ou indéterminé : /La^δKaR/ les ou des troupes (ce mot a donné en français "lascar") ce qui signifie que le graphisme consonantique du mot /L^δKR/ est discriminant du point de vue lexico-sémantique et qu'il n'y a pas nécessité de procédure syntaxique ou sémantique pour identifier le mot paradigme. Par contre, le choix déterminé-indéterminé à l'intérieur du même paradigme s'obtient par le contexte.

1.2. Correspondance non biunivoque entre [C,SV] et [C,SV,VL]

a) Ce que l'on écrit /MâLK/ {graphisme de la forme (C.VL.C.C) est du type [C,SV,VL]}, peut, après vocalisation, être lu /MâLèK/. Par contre, si l'on retire la voyelle longue /â/ du mot source /MâLK/, le graphisme MLK devient ambigu et fournit les mots suivants :

⁴ MOINFAR, M. Dj., 1987, Ch.0, p.20.

/MâLèK/	(propriétaire)
/MaLaK/	(ange)
/MoLK/	(pays)
/MèLK/	(propriété)

b) Le mot /BâND/, qui représente une séquence de la forme (C.VL.C.C) correspondant à un graphisme du type [C,SV,VL] est non ambigu. En revanche, la suppression de la voyelle longue /â/ fournit la forme /BND/ du type [C,SV] qui, après vocalisation, produit les mots suivants de la langue :

/BâND/	(bande) mot source écrit en graphisme [C,SV,VL]
/BaND/	(lac)
/BaNDi/	(prisonnier)

c) Même situation pour le mot /aWLâD/ de la forme (VL.SV.C.VL.C) qui en graphisme du type [C,SV] s'écrit /WLD/ auquel correspondent:

/aWLâD/	(enfants) mot source écrit en graphisme [C,SV,VL]
/WaLaD/	(fils de)
/WaLiD/	(né ou enfanté)
/WâLèDa/	(mère)

1.3. Graphisme et polymorphisme

Nous avons dit que le graphisme [C,SV,VL] est très souvent non ambigu. Il existe donc, cependant, des cas d'ambiguïtés c'est-à-dire des cas où à [C,SV,VL] correspondent plusieurs formes (C,SV,VL,VB1), (C,SV,VL,VB2) etc. Par exemple :

/BâR/, à partir de notre algorithme d'insertion de voyelles (voir IV), est polygraphique et produit les mots suivants :

/BâRi/	(portable)
/BâR/	(portage)

(ici, "â" dans /bâr/ est une voyelle longue et s'écrit explicitement dans le graphisme [C,SV,VL] mais elle ne s'écrit pas dans un graphisme de consonnes seules). Par contre, /BRNJ/ n'est pas polygraphique après vocalisation et produit le mot suivant : /BRNJ/ → /BèRèNJ/ qui, cependant, lui est encore polysémique et signifie riz ou laiton. Il y a donc, dans ce cas, réduction totale de la polygraphie mais pas de réduction de la polysémie.

Cet exemple montre que dans la correspondance :

	C		
	ou		
	C,SV		
graphisme	ou	→	polysémie
	C,SV,VL		

on doit introduire le concept intermédiaire de polygraphisme, c'est à dire:

	C				
	ou				
	C,SV				
graphisme	ou	→	polygraphisme ou non	→	polysémie ou non
	C,SV,VL				

Remarque.

Le persan, langue indo-européenne, n'a pas les mêmes limitations, pour ce qui est le nombre des consonnes, que les langues sémitiques dont la racine consonantique nominale est trilitère.

Exemples :

/PRMN/	→	/PiRâMùN/	(aux environs)
/PRSTDN/	→	/PaRaSTiDaN/	(adorer)
/STXRJ/	→	/èSTèXRâJ/	(extraction)
/DRMNDG/	→	/DaRMâNDaGi/	(misère)
/DRMNP≤R/	→	/DaRMâNPa≤iR/	(soignable)

D'une façon générale, l'expérience linguistique et la pragmatique du langage montrent qu'en persan, plus il y a de consonnes dans un mot et plus s'affaiblit la polysémie du graphisme [C,SV,VL] et aussi la polysémie du simple graphisme consonantique. Cela signifie que, pour l'essentiel, l'étude de la polysémie du graphisme [C,SV] peut se réduire aux formes binaires et trilitères.

Compte tenu des remarques précédentes, nous nous sommes orientés, tout d'abord, vers l'étude des graphismes [C,SV,VL] et nous avons cherché à traiter les mots les plus polysémiques, en commençant par les formes de 3 consonnes car ces dernières sont les plus fréquentes en dari.

Il faut bien souligner que le graphisme consonantique est plus polysémique que le graphisme [C,SV,VL] complet, mais il permet de donner un caractère systématique à notre étude.

3. TRANSCRIPTION ⁵

Pour la réalisation de notre travail, nous avons choisi le système de transcription suivant, en tenant compte des contraintes imposées par les claviers des ordinateurs que nous utilisons de sorte que chaque graphème ne représente qu'un seul phonème du dari. Il diffère de celui employé par les linguistes européens travaillant sur le persan.

a) La transcription des voyelles du dari

1- {ٲ}	/i/	[i]	ex.	{شیر}	/ʃiR/	(lait)
2- {ٲ}	/e/	[e]	ex.	{شیر}	/ʃeR/	(lion)
3- {ٲ}	/ü/	[u]	ex.	{دو}	/DùD/	(fumée)
4- {ٲ}	/u/	[o]	ex.	{گوش}	/Gu,ʃ/	(oreille)
5- {آ}	/â/	[a:]	ex.	{آب}	/âb/	(eau)
6- {ا}	/a/	[a]	ex.	{ابر}	/abr/	(nuage)
7- {ه}	/è/	[e]	ex.	{چله}	/JèLD/	(peau)
8- {و}	/o/	[ɔ]	ex.	{گل}	/GoL/	(fleur)

Remarque.

(1,2,3,4,5) représentent les voyelles longues et (6,7,8) celles qui sont brèves. Celles-ci à l'exception de /a/ à l'initiale ne s'écrivent pas.

⁵ AFZALI, Razeq, 1986, Ch.II, pp.34-46.

b) La transcription des consonnes du dari

1- {ب}	/B/	[b]	2- {پ}	/P/	[p]	3- {ت}	/T/	[t]
4- {ث}	/Σ/	[s]	5- {ج}	/J/	[dz]	6- {چ}	/&/	[tch]
7- {ح}	/H/	[h]	8- {خ}	/X/	[kh]	9- {>}	/D/	[d]
10- {ذ}	/≤/	[z]	11- {ز}	/R/	[r]	12- {ر}	/Z/	[z]
13- {ژ}	/#/	[j]	14- {س}	/S/	[s]	15- {ش}	/δ/	[ʃ]
16- {ص}	/μ/	[s]	17- {ض}	/£/	[z]	18- {ط}	/τ/	[t]
19- {ظ}	/δ/	[z]	20- {ع}	/·/	[ʔ]	21- {غ}	/ξ/	[gh]
22- {ف}	/F/	[f]	23- {ق}	/Q/	[qh]	24- {ک}	/K/	[k]
25- {گ}	/G/	[g]	26- {ل}	/L/	[l]	27- {م}	/M/	[m]
28- {ن}	/N/	[n]	29- {و}	/W/	[w]	30- {ه}	/θ/	[h]
31- {ی}	/Y/	[y]						

Remarques.

L'opposition des majuscules et des minuscules étant inconnue dans l'écriture du dari, nous avons utilisé les majuscules et quelques signes spécifiques pour transcrire les consonnes. De même, plus haut, les minuscules sont utilisées pour la représentation graphique des voyelles.

En dari, il y a trois catégories de phonèmes homophones non homographes :

1. (10, 12, 17, 19) qui représentent le phonème /z/.
2. (4, 14, 16) qui représentent le phonème /s/.
3. (3, 8) qui représentent le phonème /t/.

4. MÉTHODE UTILISÉE

4.1. Pour décrire la méthode prenons, par exemple, le cas d'une forme consonantique trilitère (C.C.C).

Pour bien illustrer cette méthode, nous énonçons d'abord les règles de vocalisation suivante (qui constituent les éléments de base d'un algorithme d'insertion de voyelles) : dans tous les cas, les formes suivantes sont possibles dans une expression trilitère (C.C.C) :

VC	une voyelle peut s'insérer devant n'importe quelle consonne
VCC	une voyelle peut précéder n'importe quel couple de consonnes
CV	une voyelle peut suivre n'importe quelle consonne
CVC	une voyelle s'intercale entre n'importe quelles consonnes
CVCC	une voyelle suit une consonne initiale et est suivie d'un couple de consonnes

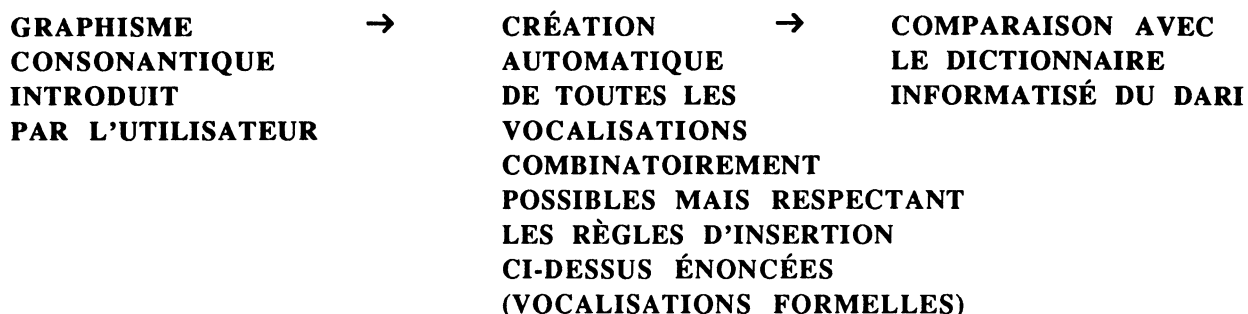
mais la forme (C.C.V.) n'est possible que si la première consonne n'est pas initiale dans le mot.

Un ensemble de trois consonnes ordonnées étant donné (par exemple MLK), on peut, à partir de cette règle d'insertion de voyelles (dans lesquelles l'hiatus⁶ est exclu), obtenir la liste complète de toutes les formes de mêmes consonnes et de voyelles différentes : par exemple MaLaK, MoLoK, MuLuK, MiLiK, MâLâK, MèLèK, MùLùK ...).

⁶ L'hiatus est absent en dari.

Parmi toutes ces formes, seules certaines constituent des mots trilitères de la langue, par exemple MaLèK, MaLèKè, MoLK, MoLKè, MoLK_i, MèLK, MèLKè, MèLK_i, MaLaK, MaLaKè, MaLaK_i, MaLèKa, MoLùK, MoLùKè, aMLâK, aMLâKè, MaLâKè, Mâlèk, Mâlèkè. Ces formes trilitères correctes sont repérées automatiquement à l'aide d'un dictionnaire informatisé du dari, élaboré indépendamment, et vocalisé ⁷. Cette procédure est évidemment renouvelée pour tous les autres triplets consonantiques et le sera ultérieurement pour toutes les formes consonantiques.

Le schéma de comparaison est toujours le suivant :



Remarque.

Cette création automatique se fait en fonction des voyelles rencontrées dans les mots rentrés dans le dictionnaire en ordre alphabétique consonantique, ce qui n'implique pas de contrainte d'ordre pour les voyelles.

4.2. Pour décrire la productivité de la procédure de vocalisation formelle, on définit alors le rapport de Productivité de Vocalisation Formelle (PVF) pour l'ensemble des triplets consonantiques formels et ce comme suit:

a) soit *NVP* = Nombre de configurations de Voyelles Productives possibles, c'est à dire donnant lieu à un mot de la langue dans un triplet de consonnes. *NVP* est obtenu automatiquement à partir du dictionnaire vocalisé que nous avons constitué manuellement. Le calcul donne *NVP*=73 ce qui signifie qu'il y a en tout 73 configurations de suites de voyelles qui apparaissent dans au moins un triplet de consonnes données. Ces configurations comprennent d'une à quatre voyelles (ci-dessous on désigne par ϕ une occurrence éventuelle non-occupée par une des 8 voyelles).

1. Exemples avec une voyelle /è/ (11 apparitions dans notre corpus) :

/èLM/	ϕ 'èL ϕ M ϕ	(science)	/FèKR/	ϕ FèK ϕ R ϕ	(pensée)
/KèRM/	ϕ KèR ϕ M ϕ	(ver)	/KèʒT/	ϕ Kèʒ ϕ T ϕ	(échec,semage)
/MèLK/	ϕ MèL ϕ K ϕ	(propriété)	/NèRX/	ϕ NèR ϕ X ϕ	(prix)
/XèBT/	ϕ XèB ϕ T ϕ	(aberration)	/JèLD/	ϕ JèL ϕ D ϕ	(peau)
/&èʒM/	ϕ &èʒ ϕ M ϕ	(oeil)	/ʒè'R/	ϕ ʒè' ϕ R ϕ	(poésie)
/ʒèRK/	ϕ ʒèR ϕ K ϕ	(polythéisme)			

⁷ Le dictionnaire comprend les possibilités suivantes :

- recherche d'un mot à partir d'un ensemble de consonnes ordonnées entrées à l'écran ; ajout et suppression d'un mot donné ; liste de n-uplets de consonnes ; possibilité de faire les statistiques suivantes : 1) nombre de mots du dictionnaire ; 2) nombre de triplets de consonnes ; 3) nombre de singletons de voyelles ; 4) nombre de doublets de voyelles ; 5) nombre de triplets de voyelles ; 6) nombre de quadruplets de voyelles ; 7) nombre de mots possibles ; 8) nombre de configurations de voyelles possibles.

2. Exemples avec deux voyelles /èi/ (5 apparitions) :

/èLMi/	ϕ'èLϕMi	(scientifique)	/KèRMi/	ϕKèRϕMi	(véreux)
/KèRiM/	ϕKèRiMϕ	(crème)	/KèϑTi/	ϕKèϑϕTi	(bateau)
/JèLDi/	ϕJèLϕDi	(dermique)			

Remarque.

Dans KèRiM et KèRMi la configuration des voyelles n'est pas la même si l'on tient compte du symbole ϕ. Dans le premier cas on a : ϕèi ϕ et dans le deuxième cas on a : ϕèϕi. De ce fait, cette configuration est comptée deux fois.

3. Exemples avec trois voyelles /oaâ/ (3 apparitions) :

/oLaMâ/	ϕ'oLaMâ	(savants)	/HoKaMâ/	ϕHoKaMâ	(philosophes)
/ϑo'aRâ/	ϕϑo'aRâ	(poètes)			

4. Exemple avec quatre voyelles /âaii/ (1 apparition) :

/âFaRiNi/	âFaRiNi	(encouragement, applaudissement)
-----------	---------	----------------------------------

b) soit NTV = Nombre Total (de singletons, doublets, triplets et quadruplets) de Voyelles insérées dans les triplets de consonnes.

Par définition, on a $NTV = 9 \times 9 \times 9 \times 9 = 6561$

(il y a, en effet, 4 places possibles dans un triplet de consonnes pour chacune des 8 voyelles du persan, auxquelles il faut ajouter pour la rigueur du calcul l'absence de voyelle matérialisée par le symbole ϕ.)

On pose par définition que le rapport PVF est égal à $\frac{NVP}{NTV}$

c'est à dire que dans le cas de la vocalisation relative à des triplets de consonnes on a :

$$PVF = \frac{NVP}{NTV} = \frac{73}{6561} = 0,111$$

5. RÉSULTATS STATISTIQUES OBTENUS EN FONCTION DES CALCULS OPÉRÉS

5.1. Degré général du dictionnaire par rapport aux configurations de voyelles

- Nombre de mots du dictionnaire intégré ($NMDI$)	: 504
- Nombre de triplets de consonnes obtenus automatiquement	: 219
- Nombre de configurations de quadruplets de voyelles	: 2
- Nombre de configurations de triplets de voyelles	: 39
- Nombre de configurations de doublets de voyelles	: 27
- Nombre de singletons de voyelles	: 5
- Nombre de mots contenant 4 voyelles	: 2
- Nombre de mots contenant 3 voyelles	: 118
- Nombre de mots contenant 2 voyelles	: 285
- Nombre de mots contenant 1 voyelle	: 99
- Nombre de mots possibles par vocalisation formelle pour chaque triplet de consonne ($9 \times 9 \times 9 \times 9$)	: 6561
- Nombre de mots possibles (NMP) pour les 219 triplets de consonnes (6561×219)	: 1436859

d'où le rapport : $\frac{NMDI}{NMP} = \frac{504}{1236859} = \frac{1}{3000}$

2 - Apparition des différentes séquences de voyelles productives au nombre de 504, selon les règles d'insertion de voyelles précisées au chapitre IV.

VOYELLE(S) ET NOMBRE D'APPARITION DE CES VOYELLES

a	59	aâ	38	ai	38
aa	35	aè	27	âè	25
o	21	êâ	19	aâè	14
aâi	13	oa	12	oè	12
è	12	èa	9	aaè	8
où	8	âa	8	oâ	8
aiè	7	oi	7	èâè	6
ââ	6	èè	6	uè	6
âèè	5	âaè	5	oâè	5
èi	5	aïi	4	oùè	4
â	4	èâi	3	âèa	3
aâa	6	oaâ	3	ââè	3
ââi	3	aia	3	èaè	3
oo	3	ia	3	u	3
oâi	2	ooi	2	ui	2
uâ	2	aaaè	1	âaïi	1
èââ	1	aai	1	oùi	1
âaa	1	âai	1	èèâ	1
èèè	1	èèi	1	èiè	1
èii	1	aèâ	1	aèè	1
aèa	1	aèi	1	aoi	1
aoè	1	âuè	1	èai	1
ooè	1	aù	1	au	1
iâ	1	iè	1	ûâ	1
ao	1				

6. CONCLUSION

Les deux rapports $NMDI/NMP$ et $PVF = NVP/NTV$ sont très faibles ce qui montre qu'il y a d'emblée une contrainte vocalique au niveau lexical solidaire du graphisme [C,SV,VL] . Donc, on ne rencontre pas n'importe quelle succession vocalique et, de plus, la succession vocalique est solidaire de la succession consonantique pour une suite de consonnes déterminée. Sur les 73 formes vocaliques productives, toutes ne sont pas possibles avec n'importe quelle suite d'éléments représentant un mot de la langue. Par exemple, on peut avoir:

/ZaDaN/	(battre)	mais pas	/ZoDaN/
/ʒoDaN/	(devenir)	mais pas	/ʒaDaN/

Cela signifie que la combinaison consonnes-voyelles est très restrictive, et que typologiquement il n'y a presque aucune forme vocalique qui, d'une façon universelle, pourrait être compatible avec toute forme (V.C.V.C.V.C.V.) et donc a fortiori [C,SV,VL]. Cela permet partiellement de comprendre pourquoi en dari on peut ne pas écrire les voyelles brèves, car le travail de reconstitution d'une forme lexicale complète pour un dariphone à partir d'une forme (C.SV.VL.) se trouve, de par les propriétés mêmes du lexique, très facilité.

On se propose, dans une étude ultérieure, d'examiner la diversité sémantique générée par un même graphisme ordonné [C,SV,VL] trilitère. Car nous l'avons vu, si le polygraphisme des formes trilitères est assez faible, cela ne préjuge pas du degré de leur polysémie. En particulier, on étudiera les traits sémantiques communs aux diverses formes vocalisées qui correspondent à un graphisme [C,SV,VL] donné ou à un graphisme consonantique donné.

BIBLIOGRAPHIE

AFZALI, Razeq, *Analyse morphosyntaxique automatique du dari (persan d'Afghanistan) et mise au point d'un système d'interrogation de bases de données textuelles en langage naturel*, Thèse de doctorat, Paris-Sorbonne, 1986, 250.

ANDREEWSKY, Alexandre, FLUHR, Christian, "Les langages naturels. Pour l'accès aux banques de données textuelles", in *Convention informatique*, Paris, septembre 1982.

CATACH, Nina, *La phonétisation automatique du français*, Paris, éd. CNRS, 1984.

SAUSSURE, Ferdinand (de), *Cahiers Ferdinand de Saussure*, Genève, Éditions universitaires, 1941.

JAKOBSON, Roman, "Principes de phonologie", appendice 1 des *Principes de phonologie* de N.S. TROUBETSKY, Paris, 1949.

JAKOBSON, Roman, WAUGH, L., *La charpente phonétique du langage*, trad. par A. Kihm, Éd. de Minuit, 1980.

LAZARD, Gilbert, *Grammaire du persan contemporain*, Paris, Klincksieck, 1957, 297.

MOINFAR, M.Djafar, "Phonologie quantitative du persan", in *Documents de Linguistique Quantitative*, 16, Paris, Jean-Favard, 1973, 256.

MOINFAR, M.Djafar, "Grammaire comparée de l'arabe et du persan, premier fascicule : grammaire de l'arabe", in *Documents de Linguistique Quantitative*, 12, Paris, Jean-Favard, 1973, 256.

MOINFAR, M.Djafar, "Grammaire comparée de l'arabe et du persan", deuxième fascicule : grammaire du persan", in *Documents de Linguistique Quantitative*, 34, Paris, Jean-Favard, 1978, 256.

ORANSKIJ, J.M., *Les langues iraniennes*, Paris, Klincksieck, 1977.

ROMAN, A, *Grammaire de l'arabe*, Que sais-je ?, Paris, Presses Universitaire de France, 1990.