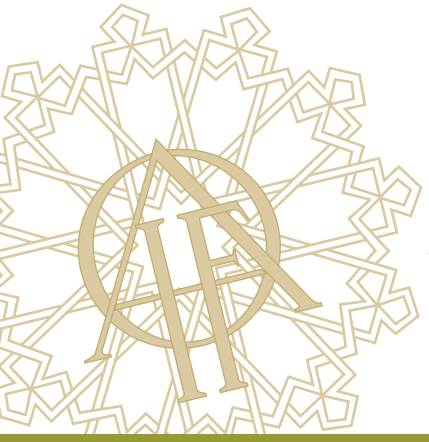

ANNALES ISLAMOLOGIQUES

en ligne en ligne en ligne en ligne en ligne en ligne en ligne en ligne en ligne en ligne

AnIsl 29 (1995), p. 283-311

Christian Gaubert

Analyse morphologique d’un texte arabe par ordinateur: méthode d’évaluation, résultats.

Christian GAUBERT

ANALYSE MORPHOLOGIQUE D'UN TEXTE ARABE PAR ORDINATEUR : MÉTHODE D'ÉVALUATION, RÉSULTATS

Les possibilités d'analyse morphologique automatique de l'arabe et d'extraction de racines mises en évidence par Claude Audebert et André Jaccarini [1] dans leurs travaux récents ouvrent de nombreuses perspectives dans le domaine du traitement automatique de textes arabes. La force de la démarche adoptée réside autant dans le respect des ambiguïtés naturelles et graphiques de la morphologie que dans l'efficacité de l'algorithmique employée. Une grammaire morphologique peut être considérée comme un objet susceptible de modifications et donc de mises au point et d'affinements : ce sont les *variations de grammaire* [2].

La morphologie du nom trilitère retient l'attention par la position prédominante de cette catégorie de mot par rapport aux autres, que ce soit les verbes, les tokens, au sens défini par Audebert et Jaccarini [3], ou les autres noms, quadrilitères, bilitères ou propres. Une première grammaire [4] non déterministe [5] évoluée, baptisée G2, a permis de reconnaître la plupart des obstacles créés par la prise en compte de l'ambiguïté, celle du modèle programmé comme celle due à la représentation graphique de la morphologie arabe. Nous proposons ici une méthode d'évaluation et de développement des grammaires morphologiques fondée sur l'analyse de textes. Cette méthode a pu être mise au point grâce à l'élaboration préalable d'un logiciel d'analyse aisément transformable en un module linguistique indépendant inclus dans un programme plus général.

1. Cette étude s'inscrit dans la suite logique des travaux publiés depuis 1987 par Claude-F. Audebert et André Jaccarini. Elle en exploite la terminologie et les résultats principaux. Nous nous contenterons donc de rappels généraux sur le fonctionnement des grammaires formelles sans en détailler les fondements théoriques.

2. Ce concept a été exposé par Audebert et Jaccarini dans « Méthode de variations de grammaire et algorithme morphologique, vers un extracteur de racine en arabe » in *BEO* XLVI, 1994, p. 77-91.

3. Cf. *idd.* « De la reconnaissance des mots outils et des token » in *AnIsl* XXIV, 1988, p. 269 n. 2.

4. Cf. *BEO* XLVI, p. 82-86.

5. Est non déterministe une grammaire qui peut présenter au cours de l'analyse du mot plusieurs possibilités qui devront être explorées systématiquement.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

1. LA GRAMMAIRE FORMELLE Gn3 : UNE DESCRIPTION SOUPLE DES NOMS À RADICAUX TRILITÈRES.

Une grammaire représentée par un automate peut être décrite comme un ensemble d'états dont un état initial E_i et un état final E_f. Les variables d'entrée sont des mots ou suites de caractères pris dans un ensemble désigné par *alphabet*. Chaque état constitue une étape de l'analyse du mot, l'étape suivante étant atteinte par la lecture d'un caractère selon une catégorie linguistique précise : ce sont là les transitions ou règles de réécriture, aisément représentables par arcs orientés dans un diagramme. Chaque règle ou arc correspond à une catégorie linguistique précise à l'intérieur du mot. Des transitions particulières appelées *epsilon-transitions* permettent le passage direct d'un état à un autre lorsque la ou les catégories intermédiaires ne figurent pas dans le mot.

Si le dernier caractère du mot analysé emprunte un arc qui peut mener directement à l'état final, le mot est accepté et donc conforme à la grammaire. Notons qu'il peut l'être pour plusieurs raisons, en vertu de l'indéterminisme du modèle. Si au contraire le dernier caractère ne peut être relié à l'état final, le mot est refusé et n'est donc pas formé suivant les règles de cette grammaire.

La grammaire non déterministe Gn3 est une version légèrement modifiée [6] de G2. Elle procède dans ce cadre à une analyse fine mais sans souci d'exhaustivité des préfixes et suffixes éventuels du nom arabe, dans le cas réduit du trilitère sain. Les objets soumis à son verdict sont des mots formés d'une suite de caractères arabes, à l'exclusion des signes de vocalisation externe *fatḥa, ḍamma, kasra, sukūn, waṣla, tanwīn et šadda* ; la *hamza* est en revanche tolérée sous toutes ses formes. Les catégories du nom arabe sont ainsi clairement mises en évidence et, une fois détectées, doivent permettre de cerner les possibilités de radicaux du nom et d'émettre, dans certains cas, des hypothèses sur la fonction du mot dans la phrase.

Le modèle Gn3(c,c,c) [7], reproduit dans la figure 1 sous la forme de graphe de transitions, constitue un des états du développement de cette grammaire ; il sera pris comme référence pour la suite de cet article.

La partie radicale de cette grammaire doit retenir toute notre attention. L'influence néfaste de la présence d'éléments agglutinés conduit à distinguer deux variétés de consonnes. Σ_r est l'ensemble des consonnes nommées « solides » car elles ne peuvent qu'appartenir à la racine, à la rare exception de certains infixes produits par la dérivation de la forme VIII [8]. Σ_f est son complémentaire dans l'ensemble des consonnes « saines » : il s'agit

6. Seules les modifications apportées à cette grammaire seront détaillées au cours de cette étude.

7. Nous utiliserons lorsqu'elle s'avère nécessaire cette notation fonctionelle. « n3 » est mis pour « nominal trilitère » et les caractères entre parenthèses font référence aux élément radicaux reconnus respectivement en R1, R2 et R3, soit toutes les consonnes saines dans le cas présent.

8. C'est le cas de « د » dans « إزدهار » et de « ط » dans « مصطنع ».

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

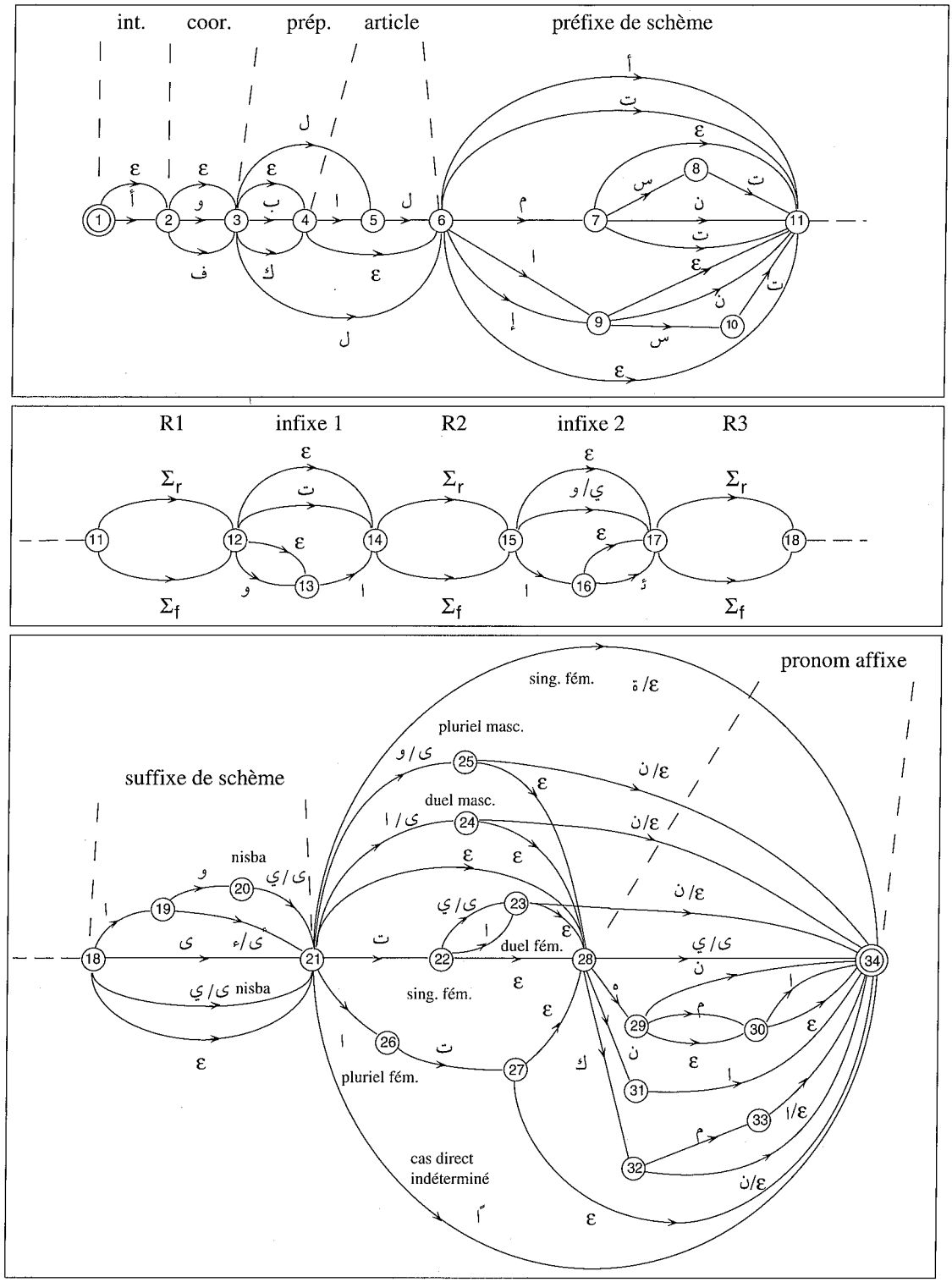

Fig. 1.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

	catégorie	abréviation	exemple	
pré-concaténation	interrogatif	int	أ	أبقلمه ؟
	coordonnant	coor	و	وغريزة
	préposition	prep	ب	بزيارة
	article	art	ال	الشيخ
racine et schème	préfixe de schème	pref	ت	التقليد
	première radicale	R1	ق	التقليد
	infixe de schème entre R1 et R2	inf1	ا	القاهرة
	seconde radicale	R2	ل	التقليد
	infixe de schème entre R2 et R3	inf2	ي	التقليد
	troisième radicale	R3	د	التقليد
	suffixe de schème	suff	اء	والغرباء
post-concaténation	nisba	nisb	ي	القرويين
	singulier féminin	sgfm	ت	بقبلاته
	pluriel masculin	plma	ين	الداخلين
	pluriel féminin	plfm	ات	كرامات
	duel masculin	duma	ان	الزائران
	duel féminin	dufm	تان	الساحرتان
	cas direct non déterminé	obnd	ا	ناقما
	postfixe	post	ها	ونسائها

Fig. 2. – Catégories de Gn3.

donc des consonnes susceptibles d'être agglomérées au nom, au titre d'article, préposition, postfixe, désinence de genre ou de nombre mais aussi d'élément du schème de dérivation.

$\Sigma_r = \{t, ğ, ḥ, ẖ, d, ḏ, r, z, š, ṣ, ḍ, ṭ, ẓ, ʿ, ġ, q\}$
$\Sigma_f = \{b, t, s, f, k, l, m, n, h\}$

Les arcs Σ_d et Σ_f en R1, R2 et R3 représentent donc les possibilités de détection par Gn3s de lettres radicales saines.

Les éléments ajoutés au mot ne faisant partie ni du schème ni de la racine seront désignés par éléments de pré-concaténation ou post-concaténation suivant leur position. L'ensemble Σ_f peut être considéré comme la réunion de deux sous-ensembles de consonne

« à risques », l'un Σ_{f1} au regard de la post-concaténation, l'autre Σ_{f3} pour la pré-concaténation.

$$\Sigma_{f1} = \{b, t, s, f, k, l, m, n\}$$
$$\Sigma_{f3} = \{t, k, m, n, h\}$$

La possibilité d'infixes composés a été ajoutée, soit « و ا » entre R1 et R2 comme dans « عواطف » et « ائ » entre R2 et R3 comme dans « ملائكة » .

De même, une nouvelle catégorie intitulée « suffixe de schème » permet la reconnaissance de schèmes tel « فعلاء » et leur comportement après l'ajout d'une nisba.

Le tableau de la (fig. 2) récapitule les catégories introduites avec un exemple d'occurrence.

Remarquons dès maintenant que la grammaire Gn3(c,c,c) peut être facilement étendue à un modèle qui permettrait la reconnaissance dans certains cas de radicaux incluant les semi-consonnes *w* ou y et les lettres contenant la *hamza*. C'est dans cette direction que nous effectuerons des variations de grammaire. Il est clair que si la racine est parfaitement saine, aucune de ces lettres ne saurait être tolérée en R1, R2 et R3.

2. Mise en œuvre informatique.

La programmation d'un analyseur de textes s'appuyant sur des grammaires non déterministes tel Gn3 a déjà été effectuée dans le cadre d'un langage de programmation interprété, le Lisp [9]. Cet analyseur tirait parti, dans une certaine mesure, de la structure naturelle de liste propre au Lisp. La réalisation d'un programme indépendant d'analyse de l'arabe pouvant être intégré à des logiciels tels les gestionnaires de bases de données ou les traitements de textes imposait une programmation au moyen d'un langage algorithmique courant aux possibilités étendues. Notre choix s'est porté sur le langage C, aujourd'hui très répandu ; il est dès lors envisageable de faire fonctionner nos programmes sur différents types de matériel sans en changer le moteur que constitue l'analyseur [10].

L'ANALYSEUR.

La difficulté essentielle de l'analyse à l'aide d'une grammaire telle que Gn3 vient de son indéterminisme : il faut pouvoir détecter et enregistrer les choix multiples qui apparaissent au cours du traitement afin de pouvoir exhiber les interprétations [11] du mot

9. Cf n. 2.

10. Soulignons toutefois qu'une adaptation de l'analyseur programmé en Lisp pourra s'avérer pertinente lors de la phase d'optimisation des programmes morphologiques.

11. Le terme « interprétation » sera précisé plus bas.

selon la grammaire utilisée. C'est une structure de donnée en « arbre » que nous avons adoptée pour développer les interprétations licites de chaque mot [12].

L'exemple d'analyse proposé ci-dessous dans la figure 3 montre chacune des ramifications de l'arbre solution dans deux cas donnant lieu respectivement à deux et trois interprétations. Elles sont rangées dans un ordre propre à l'investigation qui n'a aucune signification linguistique ou même alphabétique. Chaque caractère du mot est accompagné d'une catégorie abrégée.

فراشه

1/ف:coor/ر:R1 /ا:inf1/ش:R2 /ه:R3

2/ف:R1 /ر:R2 /ا:inf2/ش:R3 /ه:post

للتبرك

1/ل:prep/ل:art / :pref/ب:R1 /ر:R2 /ك:R3

2/ل:prep/ل:R1 / :inf1/ب:R2 /ر:R3 /ك:post

3/ل:prep/ل:art / :R1 /ب:R2 /ر:R3 /ك:post

Fig. 3. – Exemple de réponse produite par l'analyseur.

L'EXPLOITATION DE L'ANALYSEUR.

La mise au point des grammaires nominales et verbales comme des analyseurs est difficilement imaginable sans le recours au traitement systématique de textes variés et la validation des résultats obtenus. Après une première phase d'élaboration de l'analyseur à l'aide de quelques mots isolés, nous avons réalisé un logiciel expérimental baptisé « Sarfeyya » qui permet d'analyser un texte entier de quelques pages.

Le principe de variations de grammaire se traduit par la possibilité donnée à l'opérateur de composer la partie radicale de la grammaire en autorisant ou interdisant les consonnes, les lettres hamzées et les semi-consonnes.

L'analyse proprement dite est affichée ou enregistrée dans des fichiers au format neutre où sont reportées les interprétations des mots du texte, les racines et les schèmes détectés.

Il est important de remarquer que les progrès récents de la micro-informatique ont rendu fort acceptables les temps d'exécution de tels traitements. En effet, l'ordre de grandeur du temps d'analyse d'un texte de la taille de cet article est la seconde [13].

12. Un exposé détaillé de l'algorithme employé n'est pas ici notre propos. L'analyseur est donc considéré dans cette étude comme une « boîte noire » dont les entrées sont des textes et la grammaire employée, et dont la sortie est le détail de l'analyse de chaque mot détecté.

13. Ces performances sont fortement dépendantes du matériel utilisé ; leur détail relève de l'optimisation de l'analyseur et des techniques de programmation.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

3. L'ANALYSE D'UN TEXTE : COMMENT ÉVALUER UNE GRAMMAIRE ?

CLASSIFIER LES RÉPONSES.

Rappelons ici les hypothèses faites sur le mot graphique soumis à la grammaire Gn3 :

1) appartenance à la langue arabe standard, sans archaïsme ;
2) non vocalisé : si toutefois il l'est totalement ou partiellement, cette vocalisation est ignorée dans l'analyse ;
3) la hamza est correctement orthographiée, soit entre autres « أحد » et non « إحد », « إدارة » et non « أدارة » ;
4) le mot ne contient pas de fautes d'orthographe ou de frappe, notre propos n'étant pas ici la correction orthographique.

Ces hypothèses sont assez restrictives dans la mesure où bien des textes modernes maltraitent les difficultés orthographiques en les simplifiant, en particulier dans la presse.

Il est clair qu'une grammaire du type de Gn3 fonctionne comme un filtre vis-à-vis de toute suite de caractères qui lui est proposée : il peut soit refuser un mot, soit l'accepter. Un refus signifie simplement que le mot n'est pas conforme aux seules règles de la grammaire programmée ; de même, une acceptation peut être multiple mais ne pas contenir l'analyse correcte du mot. Or cette grammaire opère à ce jour sans aucun lexique. Il n'y a donc aucune confirmation de l'existence de la racine extraite. Le schème de dérivation du mot obtenu, reconstitué à partir des catégories *préfixes, infixe 1, infixe 2* et *suffixe de schème* ne subit pas non plus de contrôle d'existence : un préfixe détecté peut être incompatible avec un second infixe et l'interprétation se révéler erronée.

L'opérateur est seul juge de la pertinence des solutions. Il apparaît dès lors indispensable d'adopter une terminologie permettant de classifier les réponses en fonction de leur validité.

NOTIONS DE SILENCE ET DE BRUIT APPLIQUÉES À LA MORPHOLOGIE ARABE.

a. Terminologie.

Adoptons quelques définitions :

– Γ est un ensemble de productions écrites ou mots graphiques de l'arabe répondant à un ou plusieurs critères linguistiques : l'ensemble des noms trilitères sains, l'ensemble des verbes contenant une *hamza* en radicale, etc. Il est désigné par *champ* d'application par opposition au domaine des mots graphiques.

– G est une grammaire formelle morphologique conçue dans le but d'accepter les éléments de Γ en en détaillant les ambiguïtés et de refuser ceux qui n'en font pas partie.

Le tableau de la figure 4 permet de préciser dans une situation concrète les définitions courantes : le champ Γ est l'ensemble des noms formés à partir d'une racine trilitère saine et G est la grammaire Gn3(c,c,c).

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

ex.	Mot	Catégorie	Nr	Réponse	Racine	Schème	Classe	Cause du bruit	Conséquences	Remarques
1	اصطدام	nom sain	0	vide			silence			ط imposs.en inf1
3	رجال	nom sain	0	1/ر:R1 /ج:R2 / \:inf2/ل:R3	رجل	فعال	solution			
4	أشغال	nom sain	2	1/ أ :pref/ش:R1 /غ:R2 /:inf2/ل:R3	شغل	أفعال	solution			
				2/ أ:int/ش:R1 /غ:R2 / \:inf2/ل:R3	شغل	فعال	bruit	ambiguïté interne	B. radicalement exact	
				1/ك:prep/ن:R1 /ف:R2 /هـــ :R3 / \:duma	نفه	فعل	bruit	post-concat. : هـا	B. résiduel : نـفـه non attesté	ك en R1 est pris pour préposition
5	كنفها	nom sain	3	2/ك:prep/ن:R1 /ف:R2 /هـــ:R3 / \:obnd	نفه	فعل	bruit	post-concat.	B. résiduel	
				3/ك:R1 /ن:R2 /ف:R3 /هـــ:post/ \:post	كنف	فعل	solution bruitée			
				1/ل:prep/ل:art /ت:pref/ب:R1 /ر:R2 /ك:R3	برك	تفعل	solution bruitée			
6	للتبرك	nom sain	3	2/ل:prep/ل:R1 /ت:inf1/ب:R2 /ر:R3 /ك:post	لبر	فتعل	bruit	pré-concat. : لـل	B. résiduel : لـبـر non attesté	bruit de schème erroné
				3/ل:prep/ل:art /ت:R1 /ب:R2 /ر:R3 /ك:post	تبر	فعل	bruit	ambiguïté interne	B. radicalement attesté	incorrection : لـــل et pron. affixe
2	جدي	nom sourd	0	vide			rejet			ي imposs. en R3
7	البيت	nom concave	1	1/ \:pref/ل:R1 /ب:R2 /ي:inf2/ت:R3	لبت	فعيل	bruit sans solution	pré-concat. : ال	B. résiduel : لـبـت non attesté	
8	يخرج	verbe	0	vide			rejet			ي imposs. en préf. Sch.
9	استقر	verbe	1	1/ \:pref/س:R1 /ت:inf1/ق:R2 /ر:R3	سقر	افتعل	bruit sans solution	ambiguïté interne	B. radicalement attesté	bruit de schème erroné
10	تفطن	verbe	1	1/ت:pref/ف:R1 /ط:R2 /ن:R3	فطن	تفعل	bruit sans solution	ambiguïté interne	B. radicalement exact	
11	فعندئذ	token	0	vide			rejet			ـئـذ imposs. en suff. Sch.
12	قبل	token	1	/ق:R1 /ب:R2 /ل:R3	قبل	فعل	bruit	ambiguïté interne	B. radicalement exact	

Fig. 4. – Exemples de classement des réponses de la grammaire Gn3(c,c,c) appliquée aux noms trilitères sains et aux autres mots.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.
AnIsl en ligne

La réponse R de G à un mot arabe M quelconque est constituée de Nr éléments, Nr positif ou nul. Plusieurs cas sont alors à distinguer :

– Si Nr = 0, R est
  un *silence* si M appartient à Γ (exemple 1, fig. 4) ;
  un *rejet* si M n'appartient pas à Γ (ex. 2) ;
– Si Nr > 0, il reçoit le nom d'*ambiguïté* de la réponse : R est alors composée de Nr *interprétations* qui sont qualifiées de :
  solution si M appartient à Γ et l'interprétation est une analyse juste de M (première interprétation de l'ex. 3, soit 3.1) ;
  bruit si M appartient à G et l'interprétation est une analyse erronée (ex. 4.2) ou si M n'appartient pas à Γ (ex. 7.1).

La solution, si elle existe, est bien entendu unique : il n'est pas tenu compte des cas très rares où l'auteur jouerait sur l'écriture non vocalisée d'un mot pour lui affecter plusieurs sens.

On peut donc parler de *solution bruitée* (ex. 5.3) si R contient une solution parmi du bruit ; de même, une réponse constituée uniquement de bruit sera qualifiée de *bruit sans solution* (ex. 7.1).

b. Les causes naturelles du bruit.

L'étude des causes du bruit est capitale pour la compréhension des phénomènes d'ambiguïté engendrés par le système d'écriture graphique de l'arabe.

Un bruit, analyse erronée du mot, peut avoir plusieurs origines parfois mêlées entre elles. Pour les distinguer, nous procéderons ainsi :

– si le mot à analyser, une fois débarrassé de tout élément de post ou pré-concaténation, produit la même réponse erronée dépourvue elle aussi de ces éléments, le bruit sera considéré comme *bruit d'ambiguïté interne* (ex. 4.2, 6.3, 9.1) ;

– si le bruit persiste lorsque l'on ôte du mot à analyser les seuls éléments de post-concaténation, nous parlerons de *bruit de pré-concaténation*. C'est un phénomène de décalage de la racine : une lettre placée avant R1 est prise pour R1 et R3 elle-même est interprétée comme élément de concaténation car elle appartient à l'ensemble à risque Σ_{f3} (ex. 6.2, 7.1) ;

– réciproquement, le *bruit de post-concaténation* désigne les interprétations restées erronées malgré l'élimination des éléments de pré-concaténation [14] : les rôles sont inversés et c'est R1 qui appartient à l'ensemble à risque Σ_{f1} (ex. 5.1, 5.2).

14. Cf. l'exemple « كنفها » de la fig. 4.

Certaines causes contenues dans l'une des trois raisons majeures décrites ci-dessus peuvent faire l'objet d'un décompte particulier : c'est le cas ici du bruit engendré par la présence de l'article « ال » et de la préposition-article « لل ».

c. Les bruit propres au modèle.

Le fait que le schème de dérivation ne subisse pas de contrôle d'existence rend parfois possible des interprétations au schème erroné mais toléré par la grammaire. Ce phénomène n'est jamais une cause directe de bruit mais s'apparente plutôt à un défaut de la grammaire formelle qui pourra être éliminé par l'introduction d'une liste de schèmes. Ce bruit sera qualifié de *bruit de schème erroné* (ex. 6.2, 9.1).

Un second défaut important de ce modèle est sa tolérance vis-à-vis de la présence simultanée d'un article et d'un pronom affixe, cas impossible car il conduirait à une double détermination du nom. Ce bruit, ainsi que tout autre bruit se développant à cause d'un défaut de nature linguistique, sera noté bruit d'incorrection morphologique (ex. 6.3). Seule une modification des règles de réécriture ou une augmentation [15] de la grammaire pourra l'éliminer.

d. Les conséquences du bruit.

Si l'objectif de l'analyse est l'extraction de la racine, le bruit doit être catégorisé suivant son incidence dans cette extraction.

Il arrive en effet que parmi le bruit apparaissent des interprétations dont le radical est bien celui de la solution, que celle-ci soit par ailleurs détectée ou non : nous qualifierons ce bruit de *bruit radicalement exact* (ex. 4.2).

Au sein du bruit non radicalement exact peuvent se présenter des interprétations dont la racine est attestée, cas particulièrement embarrassant. Ces interprétations désignées par bruit radicalement attesté correspondent à l'ambiguïté même du système graphique de l'arabe non vocalisé, celle qui peut faire hésiter parfois un lecteur même averti (ex. 6.3, 9.1). L'*ambiguïté radicale* totalise le nombre de racines attestées différentes dans R.

Le *bruit résiduel* correspond aux interprétations dont la racine n'est pas attestée : le seul recours à une liste des racines attestées suffit à l'éliminer (ex. 5.1, 5.2, 6.2, 7.1).

La terminologie adoptée prend tout son sens lorsque l'on étudie la réponse des verbes et des tokens, qui ne sont donc pas éléments de ce champ Γ. Certains d'entre eux sont heureusement rejetés, mais d'autres sont acceptés et la réponse est alors clairement un bruit. Ce dernier est radicalement exact si le schème non vocalisé du mot se confond avec un schème non vocalisé nominal prévu par la grammaire. La cause de ce bruit est classée parmi les ambiguïtés internes (ex. 10.1).

15. Une grammaire est dite « augmentée » lorsque l'on sort du cadre des langages réguliers pour introduire des lexiques ou des arcs ne pouvant être empruntés que sous certaines conditions afin de modéliser des exceptions.

Ainsi, tout dépouillement de texte à l'aide d'une grammaire orientée vers l'analyse des mots et l'extraction de leur racine devra comporter la comptabilité précise de ces bruits et silences. Il est clair par ailleurs qu'une telle étude devra porter sur tous les mots du texte, sauf les noms propres et emprunts étrangers, et ceci afin de mesurer en quelque sorte le pouvoir de rejet ou « pouvoir séparateur » de la grammaire vis-à-vis des autres catégories de mots que celles pour laquelle elle est destinée.

UNE PROPRIÉTÉ IMPORTANTE DES GRAMMAIRES FORMELLES.

Une certaine catégorie de variations des grammaires s'accompagne d'une variation *linéaire* de la réponse. En effet, supposons qu'une grammaire G accepte un nombre T de transitions en deux quelconque de ses états E et F. Sa réponse R à un mot se décompose en deux réponses : celle R_1 contenant les interprétations qui empruntent les T arcs permis et celle R_2 des interprétations utilisant une autre voie (un arc epsilon-transition entre deux états situés de part et d'autre de E et F par exemple). Si l'on ajoute une transition possible supplémentaire entre E et F, la réponse de l'automate obtenu G' au même mot se décompose de la même façon entre une R'_1 et R'_2. Il est clair que R'_1 est R_1 et que R'_2 est R_2 augmentée des interprétations utilisant la possibilité nouvelle de transition entre E et F.

La portée de cette propriété est importante dans notre cas : elle signifie que pour une seule variation concernant l'ajout d'un arc entre deux états, il suffit d'étudier la réponse de l'automate auxiliaire pour lequel les transitions entre ces états sont limitées au seul arc introduit. La réponse de l'automate complet est alors la somme des réponses de l'automate original et de l'auxiliaire. Ce procédé peut être généralisé à plusieurs arcs à condition qu'ils relient toujours les deux mêmes états. Il peut aussi par extension s'appliquer à l'introduction de sous-automates entre deux états [16].

MÉTHODE DE DÉPOUILLEMENT D'UN TEXTE.

L'opérateur désireux d'évaluer la portée d'une grammaire pourra suivre naturellement ces étapes :

1) une analyse préalable consistant en l'extraction « manuelle » des racines et leur report dans une base de donnée. Celle-ci devra comporter au moins le mot, sa catégorie (nom, verbe, token et autres), sa racine et le type de cette dernière (saine, concave, etc.) ;

16. Une étude complète de ces propriétés dépasse la portée de cet article. Remarquons toutefois que le recours à la formulation algébrique (Cf. *BEO* XLVI, annexe I) permet de démontrer rapidement ce cas de linéarité.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

2) l'utilisation du logiciel pour produire l'analyse « automatique » du texte ;
3) la validation du résultat de 2 par la vérification de chaque interprétation et l'enrichissement de la base des mots conformément à la terminologie exposée ci-dessus ;
4) L'étude et la synthèse des statistiques issues de 3, au moyen par exemple de taux mesurant la réussite des extractions, la fréquence de telle cause de bruit, etc.

Les variations de grammaire ont pour but l'élimination d'une source de bruit ou la réduction des silences observés à l'issue de l'étape 4. Les tentatives de modification du champ d'étude des grammaires s'effectuent également par variations.

Ces variations impliquent la reprise des phases 2 à 4 et la comparaison des résultats obtenus en 4. Si l'une d'elles entre dans le cadre de la propriété de linéarité, il suffit d'étudier l'analyse du texte produite par la seule présence du ou des arcs nouveaux entre les états souhaités.

L'étude des origines du bruit peut amener à définir de nouvelles variables de dépouillement afin de pouvoir repérer les causes majeures et leur éventuelle interpénétration. Un dialogue s'amorce entre l'étape 3 et la suivante jusqu'à l'obtention d'une description satisfaisante des phénomènes survenant lors de l'analyse.

On peut mesurer au moyen de taux adéquats la réussite globale de la grammaire dans sa mission d'analyse morphologique. Ces taux peuvent ensuite servir d'éléments de comparaison de deux grammaires à l'épreuve d'un même texte, ou d'une même grammaire devant deux textes de nature différente tel un article de presse ou une poésie.

C'est la démarche que nous avons adoptée ici. Le dépouillement a été facilité par le fait que l'analyse d'un texte dans sa forme de fichier informatique standard peut être relue par un simple gestionnaire de base de donnée. Les fonctions de recherche et de tri permettent ensuite d'établir les statistiques souhaitées.

4. Analyse d'un texte littéraire contemporain

ÉTAPE 1 : LE DÉPOUILLEMENT « MANUEL ».

C'est sur un texte littéraire contemporain que nous avons procédé à une première analyse nominale. Il s'agit des deux premiers chapitres de la nouvelle de Yaḥyā Ḥaqqī intitulée « Qindīl Umm Hāšem » [17] dont l'intégralité est reproduite en annexe. Ce texte répond aux hypothèses exposées ci-dessus mais comporte naturellement de nombreux noms propres et emprunts de dialecte cairote. Les dialogues, qui représentent une part négligeable de la masse des mots, ont été coupés. Cependant une comptabilité précise

17. *Mu'allafāt Yaḥyā Ḥaqqī, al-qiṣaṣ 1*, al-Hay'a al-miṣriyya al-ʿāmma li-l-kitāb, Le Caire, 1990, p. 59-69.

s'imposait pour mesurer l'ampleur réelle de ce qui sera toujours l'ennemi du traitement automatique d'un texte.

Le tableau I (fig. 5) distingue pour chaque catégorie de mots le nombre des premières occurrences du nombre total d'apparitions dans le texte. Dans le premier cas comme dans le second *a fortiori* les suites de caractères telles « جدى » et « لجدى » sont comptées pour deux mots distincts. Il s'avère ainsi, et ceci ne doit pas nous surprendre, que pour un texte d'un total de 1300 mots environ près de la moitié est constituée de noms trilitères, 30 % de tokens et moins de 20 % de verbes trilitères. Les noms propres et toutes les autres catégories peuvent être pris pour quantité négligeable, leur total étant circonscrit autour de 5 %. La priorité donnée à la mise au point de grammaires nominales apparaît dès lors pleinement justifiée.

Le tableau II donne le résultat du dépouillement radical « manuel » de chaque nom trilitère. Les abréviations adoptées sont les suivantes :

« c » désigne une consonne élément de Σ_d ou Σ_f ;
« s » désigne une semi-consonne *w* ou *y* ;
« h » désigne la hamza ;
« = » en troisième position signifie que la racine est redoublée : R3 = R2.

Les noms sourds à consonnes solides sont distingués suivant deux catégories : la forme dissimilée « مرور », notée *cc=2*, et la forme assimilée « مرة », notée *cc=1*.

On peut constater la forte proportion des racines saines *ccc*, qui totalisent près de la moitié des cas, et donc 22 % du texte total. Les autres cas les plus fréquents sont les radicaux concaves *csc*, puis les sourds *cc=*, défectueux *ccs*, assimilés *scc* et à première radicale hamzée *hcc*. Les cas présentant une double irrégularité ne totalisent que 3 % de la masse des racines nominales trilitères de ce texte.

L'étude similaire pour les verbes trilitères présentée dans le tableau III montre que là encore, les racines sont saines pour presque la moitié des cas, concaves pour le quart tandis que les autres irrégularités se présentent avec une fréquence comparable à celle du nom.

Ces faits sont à comparer aux statistiques portant sur les 4814 racines trilitères attestées [18] d'un dictionnaire arabe classique : la proportion de racines trilitères saines s'élève à 62 %. La suite des autres irrégularités rangées par fréquence décroissante respecte le même ordre que celui issu de notre texte.

L'enjeu de cette recherche commence ainsi à se dessiner : doit-on mettre au point autant de grammaires qu'il y a de cas d'irrégularités dans la morphologie – qu'elle soit

18. D'après une étude menée par ʿAlī Ḥilmī Mūsā à partir du dictionnaire *Aṣ-ṣiḥāḥ* d'Al-Ǧawharī. « Dirāsat 'iḥṣā'iiya liǧuḏūr mufradāt al-lūġa al-ʿarabīyya », Université de Koweit, 1971. Cette étude est complétée par les travaux de Hussein Habaili, « Phonologie et morphologie de l'arabe », thèse de doctorat de 3ème cycle, Université de Paris III.

catégorie	type racine	1ère occ	répétitions	total	part	cumul
nom	trilitère	540	60	600	46,7	
	bilitère	10	2	12	0,9	
	quadrilitère	12		12	0,9	48,5
	propre	20	14	34	2,6	
	dialecte	16		16	1,2	3,9
verbe	trilitère	221	13	234	18,2	
	quadrilitère	2		2	0,2	18,4
token	trilitère	37	40	77	6,0	
	autres	88	211	299	23,3	29,2
totaux		**946**	**340**	**1286**	100,0	

Tableau I. Répartition des catégories de mots dans le texte.

type de racine trilitère		1ère occ	répétitions	total	part	cumul
sain	ccc	274	10	284	47,3	
sourd	cc=2 dissim.	26		26	4,3	
	cc=1 assim.	25	7	32	5,3	
une semi-consonne	csc	89	19	108	18,0	
	ccs	42	6	48	8,0	
	scc	25	3	28	4,7	
une hamzée	hcc	20	6	26	4,3	
	cch	6		6	1,0	
	chc	4		4	0,7	93,7
deux irrégularités	cs=	6	3	9	1,5	
	ssc	5	1	6	1,0	
	csh	5		5	0,8	
	hc=	5	5	10	1,7	
	css	3		3	0,5	
	hsc	2		2	0,3	
	scs	1		1	0,2	
	sc=	1		1	0,2	
	hcs	1		1	0,2	6,3
		540	**60**	**600**	100,0	

Tableau II. Répartition des types de racines des noms trilitères.

type racine	1ère occ	répétition	total	part	cumul
ccc	102	2	104	44,4	
cc=2 dissim.					
cc=1 assim.	14		14	6,0	
csc	48	7	55	23,5	
ccs	22	3	25	10,7	
scc	12		12	5,1	
hcc					
cch	11		11	4,7	
chc	1		1	0,4	94,9
css	5		5	2,1	
chs	2	1	3	1,3	
csh	2		2	0,9	
hss	1		1	0,4	
hcs	1		1	0,4	5,1
	221	**13**	**234**	100,0	

Tableau III. Répartition des types de racines des verbes trilitères.

Fig. 5. – Résultats de l'étape de dépouillement du texte étudié.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

d'ailleurs nominale ou verbale – ou doit-on au contraire tenter de rassembler, au prix sans doute de compromis, les cas embarrassants en un seul modèle ? Quels sont les cas réellement « embarrassants » vis-à-vis d'un traitement automatique ? Nous tenterons de répondre en montrant jusqu'à quel point le concept de variations de grammaire peut s'adapter à des problèmes précis de traitement de l'arabe.

PREMIÈRE ANALYSE : LA GRAMMAIRE DU NOM TRILITÈRE SAIN Gn3(c,c,c).

Chaque analyse se présentant sous forme d'une liste d'interprétations pour l'intégralité des mots du texte, nous nous contenterons d'en reproduire en annexe un extrait significatif. Les réponses vides, silences ou refus, n'apparaissent pas dans les listes d'interprétations. Les tableaux A de résultats montrent la répartition des réponses en termes de silence, rejet, succès et bruit concernant les noms et verbes trilitères ainsi que les tokens : c'est donc l'aboutissement de l'étape 3.

a. Gn3(c,c,c) et les noms trilitères sains.

La solution est toujours trouvée, à l'exception d'un silence dû à l'ignorance du phénomène de transformation d'infixe de la forme VIII [19]. Cette valeur minimale du silence est le résultat de l'affinement du modèle d'origine G2 par la présente méthode.

Le bruit détecté provient en grande partie d'éléments de post-concaténation qui ne sont pas des pronoms affixes, tels « ا », « ين », « ى » après R3, ce dernier cas pouvant donner lieu à cinq interprétations (cf. l'analyse de « شحاذى » montrée dans l'extrait). L'influence de la présence du « ال » est presque nulle. La plupart des cas d'ambiguïté interne sont le fait du « أ » avant R1 dont la catégorie est toujours *préfixe de schème* dans ce texte mais qui peut s'interpréter, rarement il est vrai, comme une particule interrogative. Le bruit le plus embarrassant, le bruit radicalement attesté, s'avère ici très faible.

b. Gn3(c,c,c) et les autres mots.

Le cas des noms à radical sourd mais dissimilé est sans surprise car il constitue une sous catégorie des trilitères sains : il présente donc des résultats tout à fait comparables.

Tous les autres noms devraient être rejetés par Gn3(c,c,c). Ce n'est pourtant pas le cas, plus d'un tiers des analyses donnant lieu à du bruit. Celui-ci est essentiellement dû à la présence du « ال » qui trouble fortement les réponses des *csc* et des *ccs*. Cette source de bruit est néanmoins contrôlable car elle produit des racines parasites à première radicale « ل » rarement attestées. Les éléments de post-concaténation, et à leur tête les pronoms

19. Cf. n. 8. Ce phénomène peut être modélisé mais nécessite une augmentation de la grammaire pour ne pas engendrer de bruit.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

Tableaux A. – Dépouillement des réponses de Gn3(c,c,c).

				classe				type bruit				causes...				évaluation	
NOMS/Gn3(c,c,c)	type racine	total occurrences	total réponses (Nr)	silence	rejet	solution	bruit	bruit rad. exact	bruit rad. attesté	bruit résiduel	schème inexact	pré-concat. (sauf article)	ال et لـ	post-concat.	ambiguïté interne	succès extraction	taux bruit attesté
sains	**ccc**	284	340	1		283	57	41	6	10	9		4	36	14	94,7	1,8
sourds	**cc=2**	26	31			26	5	5						5		100,0	0
	cc=1	32	31		11		31		7	24	9	7	10	16	3	0	22,6
	csc	108	53		58		53		15	38	38	5	34	11	3	0	28,3
une radicale S	**ccs**	48	34		28		34		16	18	24		24	6		0	47,1
	scc	28	10		21		10		7	3	4		2	2	4	0	-
	hcc	26	6		22		6		1	5	1			4	1	0	-
une radicale H	**chc**	4	0		4											0	-
	cch	6	0		6											0	-
	ssc	6	0		6											0	-
	csh	5	0		5											0	-
	hc=	10	0		10											0	-
	css	3	1		2		1		1			1				0	-
reste	**cs=**	9	0		9											0	-
	hsc	2	0		2											0	-
	scs	1	0		1											0	-
	sc=	1	0		1											0	-
	hcs	1	2				2			2	2			1		0	-
totaux		**600**	**508**	**1**	**186**	**309**	**199**	**46**	**53**	**100**	**87**	**13**	**74**	**81**	**25**		10,43
sauf ccc et cc=2		290	137	0	186	0	137	0	47	90	78	13	70	40	11		34,31

VERBES/Gn3(c,c,c)	type racine	total occurrences	total réponses (Nr)	silence	rejet	solution	bruit	bruit rad. exact	bruit rad. attesté	bruit résiduel	schème inexact	pré-concat. (sauf article)	ال et لـ	post-concat.	ambiguïté interne	succès extraction	taux bruit attesté
sains	**ccc**	104	59		53		59	50	3	6	5			2	49	48,1	5,08
sourds	**cc=2**	0															
	cc=1	14	11		5		11		3	8	1			2	8	0	27,3
	csc	55	25		33		25		2	23	2	1			24	0	8
une radicale S	**ccs**	25	25		13		25			25	2				25	0	0
	scc	12	4		8		4		2	2	1				4	0	-
	hcc	0															-
une radicale H	**chc**	1	0		1											0	-
	cch	11	0		11											0	-
	css	5	0		5											0	-
	chs	3	0		3											0	-
reste	**csh**	2	0		2											0	-
	hss	1	0		1											0	-
	hcs	1	0		1											0	-
totaux		**234**	**124**	**0**	**136**	**0**	**124**	**50**	**10**	**64**	**11**	**1**	**0**	**4**	**110**		8,06
sauf ccc et cc=2		130	65	0	83	0	65	0	7	58	6	1	0	2	61		10,8

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

				classe				type bruit				causes...				évaluation	
TOKENS/Gn3(c,c,c)	type racine	total occurrences	total réponses (Nr)	silence	rejet	solution	bruit	bruit rad. exact	bruit rad. attesté	bruit résiduel	schème inexact	pré-concat. (sauf article)	ال et لـل	post-concat.	ambiguïté interne	succès extraction	taux bruit attesté
sains	**ccc**	17	11		6		11	11			1			2	11	64,7	0
sourds	**cc=2**																
	cc=1	20	17		11		17		9	8		4		12	9	0	52,9
	csc	10	1		9		1			1		1			1	0	-
S	**ccs**	21	3		19		3		1	2				1	3	0	33,3
	scc	1	0		1											0	-
	hcc	4	4		2		4			4				4	4	0	-
H	**chc**																-
	cch	1	0		1											0	-
reste trilitères	**hsc**	3	0		3											0	-
non trilitères		299	34		273		34		12	22		2		19	32	0	35,3
total		**376**	**70**	**0**	**325**	**0**	**70**	**11**	**22**	**37**	**1**	**7**	**0**	**38**	**60**		31,4
ni sain ni sourd		359	59	0	319	0	59	0	22	37	0	7	0	36	49		122

affixes tels « ه »,« ها »,« ك », produisent un fort bruit radicalement attesté pour les noms à racine csc et les sourds assimilés. Les autres catégories de racines sont clairement rejetées, la taille réduite des occurrences et du bruit rendent d'ailleurs les statistiques peu significatives.

Le rejet attendu des verbes n'est pas flagrant : 40 % d'entre eux admettent des interprétations qui sont donc clairement des bruits. Il s'agit essentiellement de bruit radicalement exact dû à la coïncidence de schèmes de dérivation des racines saines décrite plus haut. Les schèmes les plus fréquemment pris pour des schèmes nominaux sont classés par ordre de fréquence dans le tableau ci-dessous.

Schèmes dont la forme non vocalisée est commune aux noms et aux verbes

Forme non vocalisée	فعل	تفعل	أفعل	فاعل
Exemple nominal	فَعْل	تَفَعُّل	أَفْعَل	فَاعِل
Exemple verbal	فَعَلَ	تَفْعَلُ	أَفْعَلَ	فَاعَلَ

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

Le rejet des verbes *csc, ccs* et *cc=1* n'est pas plus franc. Il s'explique par la présence fréquente du « ت » suffixé au verbe pour marquer l'accompli, lequel peut être pris pour une radicale R3. Il s'agit toutefois principalement de bruit résiduel.

Les tokens ne sont pas non plus totalement rejetés. En effet, nombre d'entre eux sont formés sur une racine trilitère saine et selon le schème ambigu « فعل ». Les racines sourdes de forme assimilée sont bien représentées par tokens issus de « كل » et subissent comme les noms et les verbes la forte influence de la post-concaténation. Les tokens non trilitères sont assez nettement rejetés car souvent constitués de deux lettres seulement. Ils se comportent comme les sourds assimilés dès qu'ils sont concaténés.

Parmi les bruits propres au modèle, le bruit de schème erroné est très important : il totalise 14 % des réponses. Le bruit d'incorrection morphologique s'avère en revanche très marginal : seuls trois cas se sont produits sur l'ensemble du texte.

c. Quelques mesures de la validité de l'analyse.

Une évaluation de la réussite de l'extraction de la racine peut s'effectuer au moyen d'un taux $\tau\surd$ défini ainsi :

$\tau\surd$ = (solutions accompagnées éventuellement de bruit radicalement exact)
/nombre de noms trilitères étudiés

L'influence du bruit le plus embarrassant, le bruit radicalement attesté, peut être mesurée par τ_{bra} :

τ_{bra} = bruit radicalement attesté/total des réponses

Le calcul de ces deux taux pour chacun des types de racine montre la grande capacité de Gn3(c,c,c) à extraire les racines saines de ce texte, mais aussi sa perplexité devant les autres cas qu'elle ne parvient pas vraiment à refuser, puisque τ_{bra} varie alors de 20 à 50 %. Ces taux se réduisent à 15 % si l'on suppose un contrôle de l'existence du schème, qui devrait contrer efficacement le bruit dû à l'article.

d. Un exemple d'application : la recherche par racine.

Le fait que la racine saine soit détectée, même parmi du bruit, rend possible la réalisation d'une fonction de recherche des mots par leur racine, autrement dit une indexation restreinte aux cas sains d'un texte par les racines. Se donnant une racine saine attestée $c_1c_2c_3$, l'opérateur désire trouver tous les mots du texte formés à partir de celle-ci. Il peut composer la grammaire Gn3(c_1,c_2,c_3) pour laquelle les seuls arcs permis en R1, R2, R3 sont respectivement c_1, c_2, c_3. Les réponses non vides de cette analyse sont en effet toutes les interprétations fondées sur cette racine, et un calcul simple de probabilité montre que pour un texte de taille raisonnable il est très peu probable qu'un bruit radicalement attesté vienne troubler cette recherche.

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

EXTENSION DE Gn3 À LA MORPHOLOGIE NOMINALE NON SAINE.

Les cas de racines ne présentant qu'une seule irrégularité forment la majeure partie de la morphologie non saine. Les comportements de ces racines se manifestent de deux manières distinctes :

– soit sous une forme saine où la lettre défectueuse s'insère telle une consonne radicale dans le mot ;

– soit sous une forme irrégulière où la lettre revêt une autre forme ou disparaît par assimilation ou élision.

Remarquons que dans le premier cas la grammaire Gn3 serait capable de reconnaître le mot par simple adjonction d'un arc autorisant les semi-consonne *w* et *y* pour les racines *cwc* et *cyc*. Il est aussi possible de reconnaître la forme que prend éventuellement la lettre radicale dans le second cas. Le cas des racines concaves *csc* retiendra ici notre attention car il est le seul dans la morphologie irrégulière nominale qui ne soit pas sujet aux assimilations et élisions.

Une précision de vocabulaire s'impose : il est important de faire maintenant la distinction entre la racine telle qu'elle est apparue jusqu'ici et sa représentation dans un mot, qu'il soit nom ou verbe. Nous parlerons donc de représentants de R2 pour le cas des racines concaves. R2 étant alors soit *w*, soit *y*, son représentant noté R'2 peut être l'une des quatre lettres suivantes : « و », « ي », « ا », « ئ ». Les différentes possibilités de représentation sont données dans le tableau suivant [20] :

représentant R'2	R2	exemples : forme I	formes dérivées
و	و	جوع	تكوين
ي	ي	أطياب جيل	مكيف
	و	ثياب جيران	مستعيذ مقيم
ا	و	باب	إعارة
	ي	عار	إمالة مهابة
ئ	و	سائق	
	ي	بائع	

20. Dans le dernier cas, la lettre « ئ » peut aussi représenter une *hamza* radicale R2 comme dans « سائل ».

Il est clair que les différents cas de représentation répondent à des conditions précises s'appuyant sur le type de dérivation. Si la simulation de ces conditions n'est pas possible actuellement sur Gn3, on peut en revanche par simple ajout des quatre représentants de R2 dans la catégorie *R2* mesurer les conséquences d'un élargissement du champ d'application de la grammaire en terme de bruit vis-à-vis des mot à racines non concaves.

Considérons la nouvelle grammaire Gn3(c,*cs*,c) ainsi obtenue : cette variation de Gn3 répond aux conditions de la propriété de linéarité. Nous pouvons en conséquence nous contenter de restreindre dans une grammaire Gn3(c,*s*,c) la catégorie *R2* aux seuls arcs nouvellement introduits pour analyser l'effet de l'introduction des représentants des semi-consonnes en seconde radicale [21].

SECONDE ANALYSE : LA GRAMMAIRE DU NOM TRILITÈRE SAIN OU CONCAVE Gn3(c,cs,c).

Les tableaux B donnent les résultats de l'analyse du texte par la grammaire réduite Gn3(c,s,c). La classification des réponses a été appliquée comme si le champ d'application était bien réduit aux seuls noms concaves : c'est pour cette raison que les réponses aux noms ccc sont classées soit comme bruits, soit comme refus bien que que la partie saine de Gn3(c,cs,c) les accepte.

Si tous les noms *csc* fournissent au moins une réponse, les trois quarts d'entre eux sont tels que le représentant R'2 est R2. On accède donc dans ce cas directement à la racine de type *cwc* ou *cyc*. Les autres cas sont principalement ceux pour lesquels le représentant est « ا » ou « ئ » mais aucun d'entre eux n'est véritablement ambigu car même si les deux possibilités *cwc* et *cyc* sont attestées, une seule des deux produit la forme détectée. Aussi avons-nous pris le parti dans ce dépouillement de ne plus nous restreindre aux seules attestations de cas de racines concaves, mais d'inclure aussi les formes I de ces racines. Il faut donc pour chaque interprétation d'un mot vérifier si elle peut provenir du *cwc* ou du *cyc* corrrespondant, et attribuer à chacun des deux cas le bruit adéquat.

Un mot tel « بيع » est ambigu car, bien que reconnu directement par la grammaire avec le bon radical R2 *y*, le recours à un lexique des formes I rappelle l'existence de « بيّع » issu de « بوع ». Ce mot est compté comme solution mais aussi comme bruit radicalement attesté bien que la réponse ne comporte qu'une seule interprétation. À l'opposé, « ساق » est un bruit car la racine n'est reconnue qu'à travers son représentant *alif*, lequel n'est produit qu'avec la racine « سوق ». Cette dernière donne lieu à un bruit radicalement exact, et l'autre possibilité « سيق » n'étant pas attestée est comptée comme bruit résiduel.

21. La linéarité de cette variation s'exprime ici simplement par Gn3(c,cs,c) = Gn3(c,c,c) + Gn3(c,s,c). Cf. n. 16.

Tableaux B. – Dépouillement des réponses de Gn3(c,s,c).

				classe				type bruit				causes...				évaluation	
NOMS/Gn3(c,c,c)	type racine	total occurrences	total réponses (Nr)	silence	rejet	solution	bruit	bruit rad. exact	bruit rad. attesté	bruit résiduel	schème inexact	pré-concat. (sauf article)	ال et لـ	post-concat.	ambiguïté interne	succès extraction	taux bruit attesté
sains	**ccc**	284	18		267		18		3	15	4				17	0	16,7
sourds	**cc=2**	26	1		25		1			1					1	0	-
	cc=1	32	2		30		2			2					2	0	-
	csc	108	125			80	45	52	3	56	1	1		10	30	74,1	2,4
une radicale S	**ccs**	48	7		44		7		2	5	1			1	6	0	-
	scc	28	11		19		11		2	9	4		4		6	0	-
	hcc	26	1		25		1			1					1	0	-
une radicale H	**chc**	4	0		4											0	-
	cch	6	0		6											0	-
	ssc	6	4		5		4			4	4		4			0	-
	csh	5	0		5											0	-
	hc=	10	0		10											0	-
	css	3	1		2		1			1				1		0	-
reste	**cs=**	9	2		7		2			2					2	0	-
	hsc	2	0		2											0	-
	scs	1	1		1		1			1		1				0	-
	sc=	1	1		1		1			1		1				0	-
	hcs	1	1		1		1			1				1		0	-
totaux		**600**	**175**	**0**	**454**	**80**	**95**	**52**	**10**	**99**	**14**	**3**	**8**	**13**	**65**		5,71
sauf csc		492	50	0	454	0	50	0	7	43	13	2	8	3	35		14,00

VERBES/Gn3(c,s,c)	type racine	total occurrences	total réponses (Nr)	silence	rejet	solution	bruit	bruit rad. exact	bruit rad. attesté	bruit résiduel	schème inexact	pré-concat. (sauf article)	ال et لـ	post-concat.	ambiguïté interne	succès extraction	taux bruit attesté
sains	**ccc**	104	2		102		2		1	3	1				1	0	-
sourds	**cc=2**	0															
	cc=1	14	0		14											0	0
	csc	55	30		29		30	27	3	30					30	49,1	10
une radicale S	**ccs**	25	0		25											0	-
	scc	12	0		12											0	0
	hcc	0															
une radicale H	**chc**	1	0		1											0	-
	cch	11	0		11											0	0
	css	5	1		4		1			1	1					0	-
	chs	3	0		3											0	-
reste	**csh**	2	0		1		1			3						0	-
	hss	1	0		1											0	-
	hcs	1	0		1											0	-
totaux		**234**	**33**	**0**	**204**	**0**	**34**	**27**	**4**	**37**	**2**	**0**	**0**	**0**	**31**		-
sauf csc		179	3	0	175	0	4	0	1	7	2	0	0	0	1		-

ANNEXE

Texte traité : les deux premiers chapitres de Qindīl Umm Hāšem de Yaḥyā Ḥaqqī

١

كان جـدى الشيخ رجب عبد اللّـه إذا قدم القاهرة وهـو صبى مع رجـال الأسرة ونسائها للتبرك بزيـارة أهل البيت ، دفعه أبوه إذا أشرفوا على مـدخل مسجد السيدة زينب ، - وغريزة التقليد تغنـى عن الدفع - فيهوى معهم على عتبتـه الرخـامية يـرشقها بقبـلاته ، وأقـدام الداخلين والخارجين تكـاد تصدم رأسـه . وإذا شاهـد فعلتهم أحد رجال الديـن المتعالين أشاح بوجهه ناقما على الزمن ، مستعيذاً باللّـه من البدع والشرك والجهالة ، أما أغلبية الشعب فتبسم لسـذاجة هؤلاء القرويين - ورائحة اللبن والطين والحلبة تفوح مـن ثيابهم - وتفهم ما فى قلـوبهم من حـرارة الشـوق والتبجيـل ، لا يجدون وسيلة للتعبير عـن عـواطفهـم إلا ما يفعلـونـه : والأعمال بالنيات . وهاجـر جدى - وهو شاب - إلى القاهرة سعيـاً للرزق . فلا عجب أن اختار لإقامتـه أقرب المساكن لجامعه المحبب . وهكـذا استقر بمنزل للأوقاف قديـم ، يواجه ميضأة المسجد الخلفية ، فى الحارة التـى كانت تسمى (حارة الميضة) . « كانت » لأن معول مصلحة التنظيم الهدام أتـى عليه من معالم القاهرة . طاش المعول وسلمت للميدان روحـه ، إنما يوفق فى المحو والإفنـاء حين تكون ضحاياه مـن حجارة وطوب ! ثم فتـح جدى متجـرا للغلال فى الميـدان أيضا . وهكـذا عاشـت الأسرة فى ركـاب « الست » وفى حماهـا : أعياد « السـت » أعيادنا ، ومواسمها مواسمنا ، ومؤذن المسجد ساعتنا .

اتسع المتجر وبورك لجدى فيه - وهذا من كرامات أم هاشم - فما كاد يرى ابنه الأكبر يتم دراسته فى الكتاب حتى جذبه إلى تجارته ليستعين به ، وأما ابنه الثانى فقـد دخل الأزهر ، واضطرب فيه سنوات وأخفق ، ثم عاد لبلدتنا ليكـون فقيههما ومأذونهما . بقى الابن الأصغر - عمـى إسماعيل آخر العنقود ، بهيئـه القدر واتساع رزق أبيـه لمستقبل أبهى وأعطـر . لعله خشـى فى مبدأ الأمـر ، عندمـا أجبره أبوه على حفـظ القرآن أن يـدفع بـه إلى الأزهر ، لأنه يرى صبية الميدان تلاحق الفتية المعممين بهذا الهتاف البذىء :

[. . .]

ولكن الشيـخ رجب سلمه ، بقلـب مفعم بالآمال ، إلى المدارس الأميريـة ، وعندئذ أعانتـه تربيته الـدينية وأصله القروى فسرعان مـا امتاز بالأدب والاتزان وتوقير معلميه ، مع حشمة وكبير صبر . إن حـرم التأنق لم تفته

النظافة . وهو فوق ذلك أكثر رجولة و أقوم لساناً وأفصح نطقاً من زملائه (المدلعين) أولاد الأفندية المبتلين بالعجمة وعجز البيان ، فما لبث أن بذ الأقران وتلألأت على سيمائه نجابة لا تخطئها العين ، فتعلقت به آمال أسرته .

أصبح ، وهو لم يزل صبياً ، لا ينادى إلا بـ (سي إسماعيل) أو إسماعيل أفندى ، ولا يعامل إلا معاملة الرجال . له أطيب ما فى الطعام والفاكهة .

إذا جلس للمذاكرة خفت صوت الأب ، وهو يتلو أوراده إلى همس يكاد يكون ذوب حنان مرتعش ، ومشت الأم على أطراف أصابعها ، حتى فاطمة النبوية – بنت عمه ، اليتيمة أبا و أما – تعلمت كيف تكف عن ثرثرتها وتسكن أمامه فى جلستها صامتة كأنها أمة وهو سيدها . تعودت أن تسهر معه كأن الدرس درسها ، تتطلع إليه بعينيها المريضتين المحمرتى الأجفان ، وأصابعها تعمل فى حركة متصلة لا تنقطع فى بعض أشغال (التريكو) من ذا الذى يقول لإسماعيل . تنبه إلى هاتين اليدين كيف دبت فيهما خلسة حياة غريبة و حساسية يقظة ، ألا تفهم ألا تفطن إلى أن دليل اقتراب عاهة العمى فى السليم هو أن تبدأ يده في الإبصار ؟

[. . .]

بين حين وآخر تحيل دمعة مترقرقة شخصه إلى شبح مبهم فتمسحها بطرف كمها وتعود إلى تطلعها . الحكمة عندها تتمثل فى كلامه إذا نطق .

يالله ! كيف تحوى الكتب كل هذه الأسرار والألغاز ؟ وكيف يقوى اللسان على الرطانة بلغة الأعاجم ؟ وكلما كبر في نظرها انكمشت أمامه وتضاءلت . قد يعلق بصره بضفيرتيها فيتريث و يبتسم . هؤلاء الفتيات ! لو يعلمن كم هى فارغة رؤوسهن !! إذا أوى إلى فراشه فعندئذ ، وعندئذ حسب ، تشعر الأسرة أن يومها قد انقضى ، وتبدأ تفكر فيما يلزمه في الغد . كل حياتها وحركاتها وقف على توفير راحته . جيل يفنى نفسه لينشأ فرد واحد من ذريته . محبة وصلت من قوتها إلى عنفوان الغريزة الحيوانية . الدجاجة القلقة ذات النظرة المتجسسة الحذرة ترقد على بيضها مشلولة الحركة ذليلة العين ، كأنها راهبة تصلى . . . هل هى هبات من فيض كرم؟ أم جزية جبار مستبد ، إرادته حديد ، له فى كل عنق طوق ، وفى كل ساق قيد؟ تعلق هذه الأسرة بولدها تعلق مسلوب الحرية والإرادة ! فأين بربك جماله ؟ جواب هذا السؤال عند قلبى . فما من مرة تمثلت فيها هذه الأيام البعيدة إلا وجدته يخفق يذكراها ، ويبدو لى وجه جدى الشيخ رجب وحواليه هالة من وضاءة ونور . أما جدتى – الست عديلة ، بسذاجتها و طيبتها ، فمن السخف أن يقال إنها من البشر ، وإلا فكيف إذا تكون الملائكة ! ما أبشع الدنيا وأبغضها لو خلت من مثل تسليمها و إيمانها .

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

٢

سنة بعد سنة و إسماعيل يفوز بالأولوية فإذا أعلنت النتيجة دارت أكواب الشربات على الجيران ، بل ربما شاركتهم المارة أيضا ، وزغردت (ماشا الله) بائعة الطعمية والبصارة وفاز الأسطى حسن - الحلاق ودكتور الحى - بحلوانه المعلوم وأطلقت الست عديلة بخورها وقامت بوفاء نذرها لأم هاشم. فهذه الأرغفة تعد وتملأ بالفول النابت وتخرج بها أم محمد تحملها فى مقطف على رأسها : ما تهل فى الميدان حتى تختطف الأرغفة ، ويختفى المقطف و تطير ملاءتها ، وترجع خجلة تتعثر فى أذيالها غاضبة ضاحكة من جشع شحاذى السيدة وتصير حادثتها فكاهة الأسرة بضعة أيام يتندرون بها. وكذلك نشأ إسماعيل فى حراسة الله ثم أم هاشم . حياته لا تخرج عن الحى والميدان ، أقصى نزهته أن يخرج إلى المنيل ليسير بجانب النهر أو يقف على الكوبرى . إذا أقبل المساء وزالت حدة الشمس وانقلبت الخطوط والانعكاسات إلى انحناءات وأوهام ، أفاق الميدان إلى نفسه وتخلص من الزوار والغرباء إذا أصخت السمع وكنت نقى الضمير فطنت إلى تنفس خفى عميق يجوب الميدان لعله سيدى العتريس بواب الست - أليس اسمه من أسماء الخدم ؟ - لعله فى مقصورته ينفض يديه وثيابه من عمل النهار ، ويجلس يتنفس الصعداء . فلو قيض لك أن تسمع هذا الشهيق والزفير فانظر عندئذ إلى القبة. لألاء من نور يطوف بها ، يضعف ويقوى كومضات مصباح يلاعبه الهواء. هذا هو قنديل أم هاشم المعلق فوق المقام. هيهات للجدران أن تحجب أضواءه. يمتلىء الميدان من جديد شيئاً فشيئاً. أشباح صفر الوجوه منهوكة القوى ، ذابلة الأعين ، يلبس كل منهم ما قدر عليه ، أو إن شئت : فما وقعت عليه يده من شيء فهو لابسه. نداءات الباعة كلها نغم حزين.

[. . .]

ما هذا الظلم الخفى الذى يشكون منه؟ وما هذا العبء الذى يجثم على الصدور جميعها؟ ومع ذلك فعلى الوجوه كلها نوع من الرضا والقناعة. ما أسهل ما ينسون! تتناول أيد كثيرة قروشاً وملاليم قليلة ليس هنا قانون ومعيار وسعر ، بل عرف وخاطر وفصال وزيادة فى الكيل أو طبة فى الميزان . . وقد يكون الكيل مدلساً والميزان مغشوشاً، كل بالبركة ، صفوف تستند إلى جدار الجامع جالسة على الأرض ، وبعضهم يتوسد الرصيف. خليط من رجال ونساء وأطفال ، لا تدرى من أين جاءوا ولا كيف سيختفون ، ثمار سقطت من شجرة الحياة فتعفنت فى كنفها . هنا مدرسة الشحاذين. حامل كيس اللقم يثقل الحمل ظهرة ينادى :

[. . .]

والشابة التى تنبت فجأة وسط الحارة عارية أو شبه عارية :

[. . .]

صوتها الصارخ يجذب الوجوه للنوافذ ، وعيناها الساحرتان تستهويان المطلات ، فتمطر عليها أكوام من الخرق ورث الثياب فى لحظة واحدة تذوب وتختفى ، فلا تدرى أطارت ، أم ابتلعتها الأرض فغارت .

وهذا بائع الدقة الأعمى الذى لا يبيعك إلا إذا بدأته السلام و أقرأك وراءه الصيغة الشرعية للبيع والشراء .

ينقضى النهار فيودع كرش الطرشجى بقية براميله ، وتترك أقدام الخراط عملها اليومى وأدواتها ، لتعود بصاحبها إلى الدار . لا يزال الترام هنا وحشاً مفترساً له فى كل يوم ضحية غريزة . يتقدم المساء ينعشه نسيم ذو دلال . تسمع من القهاوى ضحكات غضة وأخرى غليظة « حشاشى » . وإذا دلفت من الميدان إلى مدخل شارع مراسينه سمعت ضجيج السكارى فى خمارة أنسطاسى التى يلقبها أهل الحى بفكاهتهم خمارة « آنست » . يخرج منها سكير هائج يتطوح ويتعرض للمارة :

[. . .]

أشباح الميدان الحزينة المتعبة يحركها الآن نوع من البهجة والمرح ليس فى الدنيا هم . والمستقبل بيد اللّه تتقارب الوجوه بود ، وينسى الوجيع شكايته . ويبذر الرجل آخر نقوده فى الجوزة أو الكتشينة وليكن ما يكون : تقل أصوات اصطدام كفف الموازين ، وتختفى عربات اليد ، وتطفأ الشموع داخل المشنات ، عندئذ تنتهى جولة إسماعيل فى الميدان . هو خبير بكل ركن وشبر وحجر ، لا يفاجئه نداء بائع ، ولا ينبهم عليه مكانه . تلفه الجموع فيلتف معها كقطرة المطر يلقمها المحيط . صور متكررة متشابهة اعتادها فلا تجد فى روحه أقل مجاوبة لا يتطلع ولا يمل . لا يعرف الرضا ولا الغضب إنه ليس منفصلا عن الجمع حتى تتبينه عينه . من يقول له إن كل ما يسمعه ولا يفطن له من الأصوات . وكل ما نقع عليه عينه ولا يراه من الأشباح ، لها كلها مقدرة عجيبة على التسلل إلى القلب ، والنفوذ إليه خفية ، والاستقرار فيه ، والرسوب فى أعماقه فتصبح فى كل يوم قوامه . أما الآن فلا تمتاز نظرته بأية حياة . . نظرة سليمة ، كل عملها أن تبصر .

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

PHRASE EXTRAITE DU TEXTE.

فهذه الأرغفة تعد وتملأ بالفول النابت وتخرج بها أم محمد تحملها فى مقطف على رأسها : ما تهل فى الميدان حتى تختطف الأرغفة ، ويختفى المقطف و تطير ملاءتها ، وترجع خجلة تتعثر فى أذيالها غاضبة ضاحكة من جشع شحاذى السيدة وتصير حادثتها فكاهة الأسرة بضعة أيام يتندرون

PHRASE ANALYSÉE PAR Gn3(c,c,c)

N°	Mot arabe	Analyse	Racine	Schème
1	فهذه	/ف:coor/ه:R1 /ذ:R2 /ه:R3	هذه	فعل
1	فهذه	/ف:R1/ه:R2 /ذ:R3 /ه:post	فهذ	فعل
2	الأرغفة	/ ا:art /ل:art / أ:pref/ر:R1 /غ:R2 /ف:R3 /ة:sgfm	رغف	أفعلة
3	تعد	/ت:R1 /ع:R2 /د:R3	تعد	فعل
5	بالفول	/ب:prep/ ا:pref/ل:R1 /ف:R2 /و:inf2/ل:R3	لفل	افعول
6	النابت	/ ا:art /ل:art /ن:R1 / ا:inf1/ب:R2 /ت:R3	نبت	فاعل
7	وتخرج	/و:coor/ت:pref/خ:R1 /ر:R2 /ج:R3	خرج	تفعل
10	محمد	/م:pref/ح:R1 /م:R2 /د:R3	حمد	مفعل
11	تحملها	/ت:pref/ح:R1 /م:R2 /ل:R3 /ه:post/ ا:post	حمل	تفعل
13	مقطف	/م:pref/ق:R1 /ط:R2 /ف:R3	قطف	مفعل
17	تهل	/ت:R1 /ه:R2 /ل:R3	تهل	فعل
19	الميدان	/ ا:pref/ل:R1 /م:R2 /ي:inf2/د:R3 / ا:duma/ن:duma	لمد	افعيل
21	تختطف	/ت:pref/خ:R1 /ت:inf1/ط:R2 /ف:R3	خطف	تفتعل
22	الأرغفة	/ ا:art /ل:art / أ:pref/ر:R1 /غ:R2 /ف:R3 /ة:sgfm	رغف	أفعلة
24	المقطف	/ ا:art /ل:art /م:pref/ق:R1 /ط:R2 /ف:R3	قطف	مفعل
26	تطير	/ت:R1 /ط:R2 /ي:inf2/ر:R3	تطر	فعيل
28	وترجع	/و:coor/ت:pref/ر:R1 /ج:R2 /ع:R3	رجع	تفعل
29	خجلة	/خ:R1 /ج:R2 /ل:R3 /ة:sgfm	خجل	فعلة
33	غاضبة	/غ:R1 / ا:inf1/ض:R2 /ب:R3 /ة:sgfm	غضب	فاعلة
34	ضاحكة	/ض:R1 / ا:inf1/ح:R2 /ك:R3 /ة:sgfm	ضحك	فاعلة
36	جشع	/ج:R1 /ش:R2 /ع:R3	جشع	فعل
37	شحاذى	/ش:R1 /ح:R2 / ا:inf2/ذ:R3 /ى:suff	شحذ	فعالى
37	شحاذى	/ش:R1 /ح:R2 / ا:inf2/ذ:R3 /ى:duma	شحذ	فعال
37	شحاذى	/ش:R1 /ح:R2 / ا:inf3/ذ:R3 /ى:post	شحذ	فعال
37	شحاذى	/ش:R1 /ح:R2 / ا:inf2/ذ:R3 /ى:nisb	شحذ	فعال
37	شحاذى	/ش:R1 /ح:R2 / ا:inf2/ذ:R3 /ى:plma	شحذ	فعال
38	السيدة	/ ا:pref/ل:R1 /س:R2 /ي:inf2/د:R3 /ة:sgfm	لسد	افعيلة
39	وتصير	/و:coor/ت:R1 /ص:R2 /ي:inf2/ر:R3	تصر	فعيل
40	حادثتها	/ح:R1 / ا:inf1/د:R2 /ث:R3 /ت:sgfm/ه:post/ ا:post	حدث	فاعلة
41	فكاهة	/ف:R1 /ك:R2 / ا:inf2/ه:R3 /ة:sgfm	فكه	فعالة
43	بضعة	/ب:R1 /ض:R2 /ع:R3 /ة:sgfm	بضع	فعلة

AnIsl 29 (1995), p. 283-311 Christian Gaubert
Analyse morphologique d'un texte arabe par ordinateur: méthode d'évaluation, résultats.

PHRASE ANALYSÉE PAR Gn3(c,s,c)

N°	Mot arabe	Analyse	Racine	Schème
5	بالفول	/ب:prep/ ا:art/ل:art/ف:R1/و:R2/ل:R3	فول	فعل
19	الميدان	/ ا:art/ل:art/م:R1/ي:R2/د:R3 / ا:duma/ن:duma	ميد	فعل
26	تطير	/ت:pref/ط:R1/ي:R2/ر:R3	طير	تفعل
32	أذيالها	/ أ:int/ذ:R1/ي:R2/ا:inf2/ل:R3/ه:post/ ا:post	ذيل	فعال
32	أذيالها	/ أ:pref/ذ:R1/ي:R2/ا:inf2/ل:R3/ه:post/ ا:post	ذيل	أفعال
38	السيدة	/ ا:art/ل:art/س:R1/ي:R2/د:R3/ة:sgfm	سيد	فعلة
39	وتصير	/و:coor/ت:pref/ص:R1/ي:R2/ر:R3	صير	تفعل
41	فكاهة	/ف:coor/ك:R1/ا:R2/ه:R3/ة:sgfm	كاه	فعلة