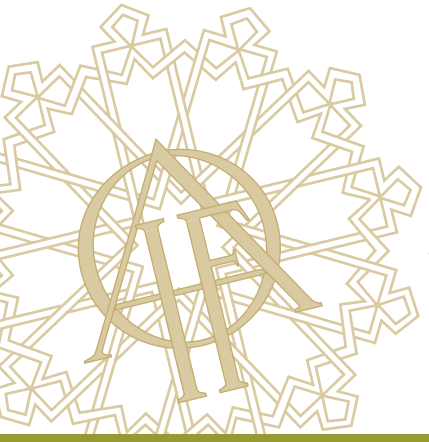

ANNALES ISLAMOLOGIQUES

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

AnIsl 44 (2010), p. 39-52

Claude Audebert

Quelques réflexions sur la fréquence et la distribution des mots-outils ou “tokens” dans les textes arabes en vue de leur caractérisation dans le cadre de l’extraction d’information.

CLAUDE AUDEBERT

Quelques réflexions sur la fréquence et la distribution des mots-outils ou *tokens*

dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information

Le terme de *token* figurant dans le titre fait allusion aux travaux menés par l'équipe du Datat[1] voici près de vingt-cinq ans. Il désigne les mots structurants que l'on ne peut réduire à des racines et qui n'obéissent donc pas à la morphologie de l'arabe. Ils recouvrent globalement ce qu'il est convenu d'appeler les mots-outils qui comprennent en outre des caractères qui s'agglutinent en début de mot et recouvrent en fait trois prépositions (*bi*, *ka*, *li*) et quelques conjonctions de coordination ou de subordination dont il sera question plus bas. Ils sont représentés graphiquement par les caractères : *b*, *k*, *l*, *w* et *f*. Ces *tokens* structurent la phrase et par là-même entraînent des attentes.

Je me propose dans cet article de faire quelques remarques de nature sémantique (domaine que nous n'avons pas abordé jusqu'alors, et, plus précisément, d'émettre quelques hypothèses sur la valeur de l'information apportée par ces *tokens*, pour ce qui est de la « caractérisation » et de la classification des textes dans une optique de recherche d'informations. Ce domaine représente aujourd'hui, vu l'ampleur acquise par le phénomène de la « numérisation » de la production écrite, une importance capitale. Ces hypothèses sur la « nature » des textes étudiés et leur possible caractérisation devront naturellement être, dans un deuxième temps, mieux formalisées et modélisées, afin de pouvoir être soumises à examen en vue de leur affinement ou même, éventuellement, leur réfutation. Je conçois donc ce travail comme une approche s'inscrivant dans une logique purement expérimentale. Sa poursuite ne pourra, par conséquent, s'accomplir que

1. Pour la partie méthodologique, cf. l'article de Jaccarini : « De l'intérêt de représenter la grammaire de l'arabe sous la forme d'une structure de machines finies».

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

grâce à des outils informatiques d'expérimentation appropriés [2]. Le présent article constitue un premier repérage d'un petit sous-ensemble de *tokens* que j'estime, en première approximation, comme plus discriminants que les autres. Ces premières tentatives de reconnaissance devraient cependant, déjà, donner lieu à des tests sur des corpus de taille raisonnable. Je pense, en effet, que les instruments de mesure dont nous disposons aujourd'hui [3] peuvent nous fournir une première indication sur la direction à suivre, une fois achevée cette première étape de « défrichage ». L'étape suivante étant d'établir une première liste de candidats de ces mots ou *associations* de mots figés que j'appellerai provisoirement « *tokens* de discours », pour mieux les distinguer des simples « *tokens* » qui ne sont que des indicateurs de structure syntaxique. J'essaie en effet de repérer les éléments qui apparaissent à la surface du texte et sont révélateurs de relations discursives pertinentes ; ces éléments pouvant par ailleurs être constitués de plusieurs formes comme les *tokens* discontinus. On l'aura donc compris, ce travail n'est qu'une amorce. Il est prévu, ensuite, de formaliser ces tentatives en ayant recours à la notion d'une « machine abstraite de simplicité maximale [4] » qui nous permettra une mise en œuvre informatique cohérente et économique. Ce dernier aspect ne sera pas abordé dans ce travail.

Nous examinerons deux textes journalistiques extraits du quotidien *Al-Ahram* du 26/1/2008, afin d'observer le comportement des *tokens* et nous demander si, à partir de leurs classes et de leurs propriétés, on peut émettre des hypothèses sur la nature des textes étudiés et leur possible caractérisation. On peut en effet poser comme hypothèse que la présence de certaines classes de *tokens* aiguillera vers certains types de textes. Cependant, dire cela, c'est poser que l'on a établi une grille de textes ce qui est loin d'être le cas, et que, inversement, cette grille même devrait être confortée, précisément, par une étude des *tokens*.

Nous dirons, soit que l'on peut, par exemple, poser qu'il existe des textes qui constituent des *comptes rendus* de faits, des *rapports* de faits ou d'événements, quel que soit le sujet ou plutôt le domaine où se passent les événements. Et d'autres où l'on discutera, interprétera, mettra en doute, récusera les mêmes événements : où l'on argumentera. On peut poser, sans grand risque, que la forme du propos ne saurait être la même dans les deux cas. Ces faits peuvent, par exemple, avoir trait à l'histoire ou à l'économie ou encore à l'art. Selon que l'on annoncera l'exposition de tel artiste ou que l'on analysera son œuvre, on peut émettre l'hypothèse que nous n'aurons vraisemblablement pas affaire aux mêmes propos et donc aux mêmes articulateurs/structurants. Dans le même ordre d'idées, on peut nous apprendre que le gouvernement a pris des mesures qui peuvent être nouvelles ou considérées comme telles dans tel ou tel domaine, et vouloir insister sur la nouveauté ou tout simplement annoncer ces mesures. Dans chacun des cas l'énoncé sera différent.

2. Cf. Audebert, Gaubert, Jaccarini, «Minimal Resources for Arabic Parsing: an Interactive Method for the Construction of Evolutive Automata», article en ligne http://www.elda.org/medar-conference/summaries/37.html,Conférence MEDAR, Le Caire, 2009.
3. Cf. *ibid*.
4. *Ibid*.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

Il serait certes insuffisant de se borner à l'analyse de deux articles de journal pour prétendre aboutir à des hypothèses cohérentes. Mais nous ne cherchons qu'à dessiner des pistes de recherche, des orientations qui sont, en fait, fondées sur une longue pratique des textes et une réflexion sur leur fonctionnement dans ce domaine. Il serait possible de décréter que certaines classes de *tokens* seront plus aptes à figurer dans une classe de *tokens* argumentatifs du fait qu'ils expriment la condition, l'éventualité ou l'irréalité, mais cela risquerait de faire passer à côté d'autres moyens linguistiques qui souvent leur sont associés. Rien n'empêche cependant d'aborder la question par les deux bouts.

L'un des textes étudiés ici tourne autour de l'annonce de nouvelles mesures d'aide aux petits revenus dans le but de réaliser la justice sociale. Qu'ils aient trait à l'économie (de manière d'ailleurs assez légère par rapport à certains articles plus techniques sur le sujet ou sur la finance), influe comme on peut le penser, sur le lexique employé. Mais influe-t-il sur les *tokens*, sur les modes de présentation des propos tenus ? Le sujet pourrait donner lieu, par exemple, à un plaidoyer pour la justice sociale et on peut imaginer qu'il présenterait des traits différents faisant intervenir des **jugements** qui modifieraient probablement moins le lexique utilisé que le mode de présentation des arguments. On s'approcherait alors des textes argumentatifs, qui restent d'ailleurs à définir. Ceux-ci pourraient avoir ou non des allures *polémiques*, être présentés sous forme de *débats*. L'on peut avoir affaire à un pur récit d'événements : guerre, catastrophe, accident. Il est rare, de toute façon, que les textes appartiennent purement à une catégorie sans comporter des éléments relevant de plusieurs classes de textes.

Nous éviterons donc de caractériser les textes à l'avance, de manière préconçue et partirons de l'examen des *tokens* présents dans les textes pour caractériser ceux-ci de manière *provisoire*. On pourrait choisir un autre point de départ, c'est celui que nous combinerons au précédent : partir de notions impliquant l'argumentation (c'est-à-dire pour simplifier, des jugements qui nous amènent d'emblée vers des moyens linguistiques ou encore l'expression d'actes de parole, que l'on peut poser au départ). Ainsi l'expression du vouloir ou du savoir entraînera des modes d'expression différents. On peut aussi examiner la mise en œuvre de moyens linguistiques propres à exprimer l'opposition entre deux idées, deux manières d'envisager deux événements, ou de notions comme la concession, la cause et la conséquence, etc.

Les *tokens*, s'ils sont privilégiés car repérables en surface, ne sont d'ailleurs pas les seuls moyens linguistiques à être mis à contribution pour l'argumentation. Nous avons déjà fait allusion, notamment, à des *tokens* de discours, mais, comme nous le verrons *infra*, les moyens linguistiques sont nombreux et variés.

Les deux textes que nous comparerons, quant à la nature des *tokens* qui y figurent, sont tirés du même numéro d'*Al-Ahram* (du 26/1/2008), mais de deux rubriques différentes. Le premier est issu de la page 1 du journal, page générale, le second de la page 11, de la rubrique intitulée *ra'y* qui le classe d'emblée dans l'expression d'une opinion.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

Texte 1 (352 mots)

Le premier texte (désormais **T1**) annonce que le gouvernement a pris de nouvelles mesures pour le retour d'une redistribution du revenu en faveur des pauvres ; c'est ainsi que le ministre du Commerce s'est adressé devant le sommet de Davos pour informer sur le fait que les hommes d'affaires ainsi que la société civile participent à une croissance juste dont tous cueilleront les fruits.

Enfin, selon un rapport britannique, l'Égypte se retrouve, au niveau mondial, au premier rang de la réforme économique.

C'est en cela que consistent les cinq lignes du titre. On en trouvera la liste des *tokens* extraite par le logiciel en ligne *Kawâkib Pro*[5], qui permet en outre le repérage des prépositions et coordonnants agglutinés en tête de mots, l'extraction des racines fréquentes, des répétitions du texte et des éventuelles citations. La liste des *tokens* sera seule commentée.

1. On note une dominante de prépositions

T1	الى	أمام	بِ	بين	ضد	على	عن	فى	لِ	مع	من	نحو
	3	2	8			7		16	14	1	6	

2. Quantifieurs

ǧamīʿ : 2
kull : 1 (*bi-kull*)

3. Conjonctions de coordination et de subordination

a. *Coordonnants*

aw : 1
fa- : 1
wāw : 30

b. *Conjonctions de subordination* :

T1	إنّ	أنْ	حتى	لأن	Relatifs
	2	3	1	1	3

– Conjonctions qui se produisent après des verbes déclaratifs (en ouverture de paragraphes)
inna : 2 (après *qāla*) et *anna* : 3 (après *aʿlana* et *akkada*)
– Conjonctions indiquant le but : *ḥattā* : 1
– Conjonctions indiquant la cause: *li-anna* : 1

4. Pronoms relatifs : *allaḏī* : 2 ; *ḥayṯu* : 1

5. Cf., dans ce même dossier, « *Kawâkib*, une application Web pour le traitement automatique de textes arabes » par l'auteur du logiciel, Christian Gaubert.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

5. Négation : 1

6. Temporel verbal : *kāna* : 1

7. Adverbe : *amsi* : 1

Remarques

1. La *dominante* écrasante *des prépositions* ne doit sans doute pas surprendre étant donné qu'elles structurent les ajouts de la phrase. Des statistiques sur leur fréquence globale et respective seraient à faire sur un corpus suffisamment vaste. Surtout en ce qui concerne les *prépositions agglutinées* dont on peut noter la grande fréquence : *bi* et *li* (14) totalisant 22 occurrences.

2. Dans ce texte, la préposition qui vient en tête est *fī* (16), suivie de *li* (14) et *bi* (8).
La préposition *li* est vouée à la fréquence sans doute du fait qu'elle sert de régime à de nombreux verbes, et cela de plus en plus fréquemment, comme *yaḥtāǧ* qui se construit avec *ilā* mais se rencontre aussi avec *li*, mais aussi du fait qu'elle permet de construire un membre de phrase où ne peut être employée l'*iḍāfa* (complément de nom).

3. Le *coordonnant wāw*, fondamental pour ce qui est de la liaison des éléments de la phrase et le lien entre unités du discours, écrase par sa *fréquence* (30), les autres mots-outils. Il serait utile de savoir tirer parti des indications d'ordre syntaxique que peuvent donner les coordonnants. On s'étonnera de la *rareté du fā'* : une seule occurrence. Quant à *aw* on notera qu'il est discriminant dans le sens où il oppose une alternative.

4. On connaît la difficulté de repérage informatique des *caractères agglutinés en débuts de mots* et qui recouvre, en ce qui concerne les *tokens*, la catégorie des *prépositions* (*bi*, *ka*, *li*) ainsi que celle des *conjonctions* de coordination et de subordination comme *wāw* et *fā'* qui appartiennent aux deux catégories, *li* qui, outre son appartenance à la classe des prépositions, est aussi une conjonction de subordination indiquant le but, le *lām* de dénégation (*al-ǧuḥūd*), et le *lām* de l'apodose (*al-ǧawāb*) dans la conditionnelle. On ajoutera le *lām* de corroboration (*lām al-tawkīd*), sans oublier le *lām* de l'impératif (*al-amr*). On ne peut en aucune façon faire abstraction de ces faits dans notre optique fondée sur les attentes qu'entraînent les *tokens* qui structurent les énoncés. La simple énumération que l'on vient de faire suggère, outre le repérage matériel des caractères, leur hiérarchisation selon la grammaire visée qui fera intervenir soit leurs pures fonctions grammaticales, soit les attentes sémantiques liées à l'extraction d'information : *li* préposition, n'aura pas la même valeur informative que le *li*, subordonnant, indiquant le but ou la dénégation par exemple.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

5. Un autre angle d'attaque serait **d'enrichir** les prépositions ou autres mots-outils en tenant compte de leur **environnement** :

Les combinaisons de *tokens*

a. *Sans séparation entre eux* : Elles sont déjà traitées comme dans MIN *bayni-hā* (classe pronoms *hā* au plur.), *min ḥayṯu*, *min ṯamma* (pas dans ce texte), *min dūni*, *ḥattā-l-āna* etc., dont une liste est déjà constituée.

BI-*kulli*, LI-*kulli*, KA-*kulli*, BI-*l-ištirâki* MAʿA, BI-*nisba* X/kaḍa, et BI-L-*nisba ilā*.

C'est à un *dictionnaire d'expressions constitué en fonction des tokens* que l'on aboutirait. Il tiendrait compte des combinaisons figées qui deviennent de véritables expressions comme *bi-ṣifati/hi* (classe des pronoms attachés) *bi-l-tālī*, *bi-l-fiʿli*, *bi-l-iḍāfati ilā* qui ne peuvent se construire qu'avec *bi*, ou encore de celles qui pourraient se construire avec deux prépositions différentes, entraînant un changement de sens comme : *min ḫilāl* « à travers » et *fī ḫilāl* « pendant » ou *bi-ǧānib* et *ilā ǧānib*. Ces quelques expressions sont en fait liées au raisonnement, ce qui en fait de véritables *tokens* de discours.

Ce type de combinaisons serait à capturer à la suite des analyses de textes opérées avec *Kawâkib Pro* et ses versions à venir, et stockées munies de leurs références aux textes analysés.

b. *À intervalles, ou tokens discontinus* : *min…ilā* (de…à)

L'importante structure *min al-… an / anna*[6] donnerait un excellent exemple des attentes de haut niveau qu'elle entraîne sur le plan syntaxique comme sur celui des notions qu'elle engage : probabilité, nécessité, obligation, souhait, jugements variés, ou au contraire, affirmation qui peut aller jusqu'à la certitude.

Laysa faqaṭ… wa lakin ayḍan ; *lam ya* (classe)*-kun… bal*, sont de très importants révélateurs pour la « caractérisation » des textes.

Tokens suivis ou précédés de lexèmes qu'ils forment avec des expressions ou des termes figés. Ex : *BI-ʿtibāri-ha* (classe /*hu*) ; *BI-faḍl*, *BI-ism* (*al-mutaḥaddiṯ BI-ism*) ; *bi-ṣifati-ha* (et sa classe), *BI-l-ištirāki MAʿA* , *BI-stimrār* ; *bi-dūni*.

Avec MIN : *al-mazīd* MIN, MIN *nāḥiya*, wa MIN *nāḥiyatin uḫrā*. En fait, *min* jouant, entre autres, le rôle de partitif sera suivi de quantifieurs (*ǧamīʿ*, *kull*, *ǧuz*).

6. L'analyse des *tokens* permet de déduire, à titre d'hypothèse, l'absence de phrases complexes, subordonnées et d'où le raisonnement ou l'argumentation sont absents (étude à vérifier sur corpus suffisamment vaste). On peut supposer que, si le texte avait tourné autour du bien-fondé de la justice sociale, par exemple, la nature des *tokens* aurait changé du fait de l'expression de jugements, de prises de position et des débats sur la question. Les remarques 5 et 6 font ressortir l'incidence immédiate des *tokens* pour l'extraction d'information et la caractérisation des textes.

6. Audebert, Gaubert, Jaccarini, « Minimal Resources ».

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

Texte 2 (592 mots)

Ce texte, classé dans la rubrique *Ra'y*, (opinions), p. 11 du même journal, *Al-Ahram*, de la même date (26/1/2008) que le texte 1, tourne autour de l'attaque israélienne de Gaza et de l'ouverture de la zone d'al-ʿArīš et Rafaḥ par le président Mubarak, afin que les Palestiniens puissent venir s'approvisionner. L'auteur du texte entend montrer que l'ouverture des frontières, décidée par le président égyptien, est humanitaire et non politique, que tous les citoyens portent les Palestiniens dans leur cœur, vérité que l'on peut tirer de la lecture de l'histoire, que l'Égypte n'abandonnera jamais les Palestiniens. Mais il met en garde ses frères contre certaines factions, d'ailleurs bien connues de l'État, contre les violences commises contre les forces de l'ordre. On sait que des échauffourées avaient d'autre part eu lieu avec l'armée et la police égyptienne, entraînant des morts. L'auteur se contente d'y faire allusion (sans les mentionner explicitement) et développe ses arguments en cinq points. Il achève son propos par une conclusion où il rappelle que la cause fondamentale de ce qui se produit à Gaza est le siège israélien et que les Palestiniens qui se sont rendus à Rafaḥ, biens conscients que le président Mubarak les a sauvés d'une famine assurée, n'ont fait qu'affirmer l'évidence historique qui prouve que l'Égypte ne les abandonnera jamais et sera à jamais à leurs côtés jusqu'à ce qu'ils recouvrent leurs droits.

Tokens du Texte 2

Ils seront présentés en regard de ceux du Texte 1 dans des tableaux afin de faciliter la comparaison.

1. Prépositions

	الى	أمام	بِ	بين	ضد	على	عن	فى	لِ	مع	من	نحو
T1	3	2	8			7		16	14	1	6	
T2	9		20	4	1	2	1	20	31	3	8	1

On note l'écrasante fréquence de *li* : 31, suivie de *bi* : 20, et de *fī* à égalité : 20; *ilā* : 9 et *min* : 8 suivent de loin. On a déjà fourni une explication de la fréquence de *li* comme suppléante de l'*iḍāfa*, dans de nombreux cas. *Bi* fournit quatre expressions figées : *bi-sabab* (à cause de), *bi-ṣadad* (face à), *bi-l-ḍabṭ* (exactement), *bi-l-fiʿl* (en fait). Pour le reste, il est le régime de verbes ou a son sens d'instrumental.

Une des difficultés du repérage informatique du *lām* est son agglutination au début des substantifs et de certains pronoms (isolés ou attachés). La seconde difficulté est ensuite de décider si le *lām* représente la préposition ou bien des conjonctions (comme dit plus haut) mais, vu sa fréquence, il doit obligatoirement être repéré. Dans le présent texte, il est tour à tour préposition, particule verbale de but (3 occurrences), de dénégation. Vocalisé en *la*, il représente la marque de l'apodose (*ǧawāb*) après *law* (1 occurrence). Comparés au Texte 1, *li* et *bi* gardent leur rang à la tête des prépositions avec une fréquence de *bi* d'un peu moins du double. *Fī* reste en très bonne place (20 contre 16 en T1).

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

2. Coordonnants

	أو	بل	ف	و
T1	1		1	30
T2	4	2	2	54

Comme en T1, le *wāw* domine tous les mots-outils de manière écrasante : 54 occurrences suivies de *aw* (4) et *bal* : (2) et *fā'* : (2). Les deux rôles de coordonnant entre propositions et de coordonnant entre paragraphes seraient à étudier. Le *wāw* n'apparaît pas ici comme conjonction de subordination.

Il est surprenant que, dans un texte argumentatif, on ne trouve que deux *fā'* qui indiquent généralement une relation temporelle ou une gradation et par suite la conséquence. Comme c'est ici le cas. *Min ṭamma* joue aussi ce rôle.

Les coordonnants *aw*: (4) et *bal* : (3) apportent un supplément d'information et sont en cela particulièrement discriminants.

3. Pronoms personnels isolés

hum : 3
hiyya : 4
huwwa : 3

Nous ne pouvons commenter ici le rôle essentiel des pronoms personnels qui peuvent marquer la séparation entre le thème et le prédicat ou remplir une de ces deux fonctions et faire office de présentatifs, et de marqueurs d'opposition au niveau discursif.

4. Pronoms ou adjectifs démonstratifs

	هذا	هذه	هؤلاء	ذلك	بذلك	كذلك
T1						
T2	8	1	1	1	1	1

5. Pronoms relatifs

	أىّ	الذى	التى	الذين	حيث	ما	من
T1		2					
T2	4	2	7	2		1	2

6. Quantifieurs

baʿḍ : 3
ǧamīʿ : 0
kull : 4

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

7. Adverbes[7]

ğamīʿan : 4
awwalan, ṯaniyan jusqu'à ḫāmisan : 1
ayḍan : 1
faqaṭ : 1

8. Noms de nombre[8]

a. Cardinaux:
waḥda-hā : 1
b. Ordinaux :
āḫir : 2 *uḫrā* : 1

9. Interrogatifs

a : *a-laysa* : 1

10. Subordonnants

	أنْ	إنَّ	أَنَّ	غير أَنّ	إلا	حتى	بِ	but لِ	لو	لو أنّ	ولو	لأنّ	عندما	Rel.
T1	3	2				1						1		3
T2	9		11	1	1	1	3	1	1	1	1	1	2	14

On note une grande variété de subordonnants dont nous avons rendu compte par le relevé des diverses phrases du Texte 2[9].

Subordonnées *conjonctives* : (20); subordonnées de *but* : (3) ; subordonnées d'*exception* : *illā* : (1) (*wa lam yakun… illā*) ; subordonnées de *concession* : *wa law liṯānia;* subordonnées *relatives* introduites par : *allaḍī, allatī, allaḍīna* ; subordonnées *temporelles* : *ʿindamā*: 2 ; phrase double : *law anna-ka… la* : (1), *law* : (1).

11. Temporels

a. Adverbes :
munḏu : 2
ʿinda'iḏin : 1
b. Particules du futur :
sa- : 1
c. Négation du futur :
lan : 2 *..abadan*

7. Voir temporels, nº 11.
8. Voir adverbes.
9. Voir *infra*, en fin d'article.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

d. Négation du passé :
lam : 2…*bal*
e. Négation du présent :
lā : 2
lā wa lā : 1
f. Interdiction :
lā + impératif.
g. Verbes :
– Indiquant le passé :
kāna: 2
– Le fait de demeurer :
ẓalla

12. Particule d'insistance
inna : 4

13. Particule restrictive
innamā : 1

Les tableaux des *tokens* présentés plus haut ne font pas ressortir un ensemble de *tokens* formant des unités de discours dans le texte. Certains *tokens* de haut niveau, c'est-à-dire entraînant des attentes syntaxiques complexes comme la conditionnelle, par exemple, devraient être pris en compte dans leur rection globale. C'est le cas de la première phrase du texte qui commence par *law anna-ka sa'alta* (si vous demandiez...) qui ne trouvera sa réponse dans l'apodose que trois lignes plus loin en *la-qāla la-ka* (il vous répondrait) tandis que se sont déroulés vingt-et-un *tokens*/mots-outils jusqu'à la fin de cette proposition. Nombre de ces phrases ont une longueur voisine, comme l'exceptive suivante qui renferme quinze *tokens* ou mots-outils ; la suivante treize. Ces faits signalent un développement et une articulation de la pensée qui devraient faire l'objet de statistiques et faire intervenir la nature des *tokens* mis en jeu afin de ne pas être confondus avec des phrases également très longues mais qui ne sont qu'un ensemble de collages d'éléments, d'énumérations, sans articulation de pensée. Une étude de la longueur des phrases apporterait en outre, de précieux renseignements.

La conditionnelle nous place d'emblée dans la **supposition**. « Si vous demandiez/Si l'on demandait ». On remarquera que cette supposition est assortie d'un ensemble de précisions – le mot de restriction serait sans doute trop fort –, amenées à circonstancier la supposition. Des expressions temporelles comme « *haḍihi-l-ayyām* », les précisions sur les acteurs (les Palestiniens), la relative qui caractérise leur action de se précipiter par vagues, leur but (afin de se procurer l'essentiel x et y), ont pour rôle de bien préciser sur quoi portera la réponse supposée de « n'importe quel citoyen égyptien », mais qui reste toutefois Égyptien. Le mouvement est ici l'inverse du premier : celui de généraliser, par l'indifférenciation du *ayy*, la réponse de *n'importe quel* et de tout citoyen de ce pays. Cet indifférenciateur sera utilisé plus bas pour insister sur le

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

fait que les Palestiniens sont les bienvenus « à n'importe quel moment » : « *fī ayyi waqt* » selon la réponse supposée, attribuée au citoyen (l'étude de cette réponse ferait ressortir l'importance de la phrase nominale de type I [10]). Disons simplement qu'elle permet l'insistance sur le pronom personnel *ka* qui incorpore le lecteur dans la discussion. L'auteur aurait pu employer à la place de *law anna-ka*, un simple *law saal-ta* qui, bien que s'adressant au lecteur, ne l'englobe pas avec autant de force. L'emploi d'une interjection qui relève du **langage affectif** et du **style direct** (à opposer à : « il vous dirait qu'ils sont les bienvenus ») serait à prendre en considération.

Un ensemble de moyens, qu'il serait trop long de détailler ici, car chacun mériterait commentaire, est mis au service de l'argumentation : pronoms personnels isolés et agglutinés, démonstratifs, adverbes de lieu ou de temps, vont inclure le lecteur dans le débat par une prise à témoin directe. Pour s'en tenir aux pronoms, une importante quantité d'informations de tout genre peut en être extraite à partir de leur présence : sur la syntaxe des phrases, le genre, le nombre, la mise en relief. Leur rôle dans la structure de la phrase nominale soit comme élément de celle-ci (*hum fī qulūbinā*), soit comme séparateurs entre thème et prédicat. Nous avons énuméré quelques-uns de ces aspects plus bas.

– Prise à témoin direct du lecteur par divers moyens : le jeu des **pronoms**, son inclusion dans la supposition : « *law anna-ka sa'alta* » par opposition à « *law sa'al-ta* ». Le *anta* l'interpelle.

– L'expression : « *wa las-nā fī ḥāğatin* » (« nous n'avons pas besoin de ») que l'on peut opposer à « il est inutile de » impersonnel (opp. *lā hāğa ilā/li, lā fā'ida fī*) fait encore intervenir les pronoms personnels. La présence du *pronom nā* dans *las-nā* fait participer le locuteur à ce groupe de citoyens qui est ainsi inclus dans le nombre de ceux qui sont forcément d'accord avec lui et ont la connaissance commune de cette histoire qui devient une preuve de ce qu'avance l'auteur : tout prouve l'aide de l'Égypte arabe à ses frères palestiniens par le passé (répété un peu plus bas : *munḏu bidāyat qadiyyati-him*, et *munḏu ʿašarāt al-sinīn* et par un tour [11] *mā kānat li… abadan*) et qui deviendra plus bas, grâce à *abadan*, indéfectible. La connaissance commune de l'histoire qu'il suffit de lire, permet d'affirmer que le comportement actuel du chef de l'État n'est que la suite logique d'un comportement historiquement identique qui ne s'est jamais démenti.

– Les *pronoms hum / nā* : (*eux / nous*), au lieu d'opposer, unissent, au contraire, Égyptiens et Palestiniens dans la même cause.

– Rôle des adverbes, *hunā*, *al-āna*, et des démonstratifs *hāḏa*, *haḏi hi-l-ayyām*, véritables *déictiques* temporels ou de lieu, ont ici aussi la fonction d'interpeller le lecteur, de le prendre à témoin et d'engager ainsi avec lui un dialogue fictif dans le présent. *Hunā : wa las-nā fī ḥāğa huna, wa huna fa-inna ʿalay-nā an*. Les démonstratifs *hāḏihi-l-ayyām*, *li-hāḏa-l-tārīḫ* indiquent la répétition et l'implicite : une histoire déjà connue et partagée par les citoyens (*li-hāḏa-l-mawqif*).

10. Phrase dont le thème est déterminé.
11. Cf. nº 7.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

– L'étude **syntaxique** révèle bon nombre de *propositions subordonnées* introduites par des *tokens* de haut niveau.

1. Phrases doubles (conditionnelles) exprimant la **supposition** :
   – *law anna-ka….la* ;
   – *ġayra anna…….iḍa mā* (*ẓahara*)
   – *fa-ʿinda'iḍin nakūnu qad* … On remarquera que l'adverbe *ʿinda'iḍin* n'est pas à proprement parler un *token* introduisant une phrase double mais que joint à *nakūnu qad*, il fait que l'ensemble exprime l'éventualité comme en français, par exemple, « *au cas où* » exprime l'éventualité.

2. Exceptive : *wa lam yakun* (*qarār ra'īs Miṣr*)… *illā*…

3. Subordonnées conjonctives exprimant l'**obligation** :
   – *Wa huna fa-inna ʿalaynā an*
   – … *ʿalay-nā* ǧamīʿan *an* naḥtarima-hu
   – allatī *yaǧibu ʿalā* ayyi (qādimin ilay-hā) *an*… *lā an*… *ḍidda-hā*… *aw*
   – wa *min ṭamma la yasiḥḥu* abadan *an* (uhāǧima-hum aw…)

   La **possibilité** :
   – wa Miṣr …*lā yumkin an* tasmaḥa *bi-an*…

4. Subordonnées *conjonctives* introduites par *anna*, qui souvent expriment une affirmation
   après 'premièrement' : *anna-l-mawḍūʿ*… *lam yakun*… *bal*
   après 'deuxièmement' : *anna* Miṣr allatī… hiyya… dawla la-hā
   après 'troisièmement' : *anna* riǧāl al-amn (thème)… *laysa fī*… *waḍa-hā bal fī kull*… muhimmatuhum… li-l-hifāẓ ʿalā (prédicat)…

5. Subordonnées conjonctives exprimant le **but** :
   Wa lasna fī ḥāǧa… hunā… li… *li-naʿrifa*…

6. Subordonnée exprimant la **concession** :
   Wa law li-ṭāniya

7. Subordonnée par le *lām de dénégation* (*lām al-ǧuḥūd*) exprimant qu'un tel n'était pas homme à faire ceci ou cela et qui introduit la négation d'une possibilité :
   Wa sa-yudriku ayyu qāri' li-haḍa-l-tārīḫ anna Miṣr *ma kānat li-*tataaḫḫar *abadan* : N'importe quel lecteur de cette histoire comprendra que l'Égypte n'était pas (un pays) à traîner les pieds, en aucune façon.

8. Subordonnée **temporelle** : wa *ʿindamā* qāla-l-ra'īs Mubārak inna…

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

9. Subordonnée **relative** :

– *anna* Miṣr *allatī* daḫaltumū-ha āminīn

– Haḏa wāǧibu-hum *allaḏī* ʿalay-nā ǧamīʿan an naḥtarima-hu

– anna Miṣr *allatī* daḫaltumū-ha āminīn… hiyya dawla la-ha qawānīnu-hā wa anzimatu-hā (sub. relative indéfinie) *allatī* yaǧibu *ʿalā ayyi* (qādimin ilay-hā) *an*…

– anna baʿḍa-l-ǧihāt wallatī hiya… ǧayyidan (tuḥāwilu)… wa (tasʿā)…

10. *ġayra anna* locution de transition, exprime l'**opposition.** Dans le troisième paragraphe, le ton change et va passer graduellement de l'explication fraternelle à l'avertissement puis à l'affirmation de cinq points « *awwalan anna* : » ; « *ṯāniyan : anna Miṣr* » ; « *ṯumma ṯāliṯan anna* » ; « *wa rābiʿan* » ; « *wa yabqā ḫāmisan anna* ». On remarquera la même structure qui dépend de la même formulation : « *wa ʿalaynā an nūrida baʿḍa-l-niqāṭ* » (« ici il nous faut produire quelques points »). Ces points seront suivis d'une conclusion.

La **négation** sous de nombreuses formes serait également à étudier dans le détail. Elle joue en effet un grand rôle dans l'expression de l'**opposition** et les négations assorties des temps qu'elles peuvent exprimer (*lam, lan, laysa* et *lā*) confirment ce rôle avec des renforcements adverbiaux du type de *abadan*, ou encore de *bal* qui, en phrase affirmative, « indique une rectification qui complète ce que l'on vient de dire: bien mieux, plus exactement [12] » et en phrase négative « exprime une rectification avec idée d'opposition [13] » *lakin* est à verser au même dossier.

lasnā fī ḥāǧa
laysa faqaṭ
lāsamaḥa-l-āh
al-mawḍūʿ lam yakun… bal
lan asmaḥ bi-taǧwiʿ… abadan
lā an yuraddida-l-šiʿārāt ḍidda-hā aw…
a-laysa kaḏālik
anna riǧāl… laysa fī Miṣr waḥda-hā bal fī kull…
lā yaṣiḥḥu abadan an… aw… lā bi-l-lafẓ wa lā bi-ayyi šay'in āḫar
… lā yumkin an tasmaḥa bi-an
fa-hāḏa lā yalīqu

La nature des phrases (nominales *vs.* verbales), l'étude des temps et des modes seraient révélateurs des structures argumentatives ou pour le moins discursives des textes.

En guise de conclusion on peut apporter les remarques suivantes :

L'argumentation qui marque ce texte (T2) *ressort de l'emploi des tokens* manifestés *en surface*. Nous n'avons pas, à dessein, fait un relevé des *racines* car celles-ci, à elles seules ne révéleraient

12. Blachère et Gaudefroy-Demombynes, *Grammaire*, p. 478.
13. *Ibid.*

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.

pas le caractère argumentatif ou non argumentatif du texte. À ce titre nous nous proposons dans un travail ultérieur, de confronter notre texte[14] avec un vaste éventail d'autres textes et, par un va-et-vient, dégager des traits susceptibles de fournir des critères en vue de leur caractérisation.

Il semble que, plutôt que les racines, dont certaines seulement peuvent être révélatrices du sujet traité, ce sont les expressions figées ou semi figées, ou locutions conjonctives, qui doivent attirer l'attention comme par exemple, en français : *au cas où*, *en admettant que*, *pourvu que*, *si tant est que*, qui introduisent condition ou supposition. Le rôle de l'adverbe de temps *ʿinda'iḍin* et les jeux possibles avec le verbe *kāna* signalé plus haut, relève d'un même type de moyen linguistique.

Les relations temporelles sont très bien représentées par divers types de *tokens* : conjonctions de subordination, adverbes, verbes indiquant le temps de manière spécifique, particules verbales, etc. Ces relations temporelles servent ici à exprimer la continuité de l'aide égyptienne par le passé et dans l'avenir et font ainsi partie de l'argumentation. Dans un autre texte, ces relations pourraient trouver d'autres emplois en vue de l'argumentation. À partir d'un balayage de textes jugés argumentatifs, de l'étude des *tokens* estimés discriminants, de la comparaison de spectres de *tokens* comparables, les remarques faites dans cet article doivent être affinées et aboutir, par l'expérimentation, à une méthode d'extraction d'information.

L'argumentation, on le voit, met en jeu des *tokens puissamment révélateurs* des caractéristiques de ce texte, mais ils sont loin d'être les seuls moyens linguistiques utilisés pour ce faire. En effet, nous avons signalé le rôle des structures linguistiques qui ne sont pas nécessairement marquées par des *tokens* comme la phrase nominale par exemple, propre à exprimer un grand nombre de catégories. Elle devrait, à cet égard, être comparée à la phrase verbale afin de faire ressortir le domaine de chacune. Il faudrait également faire rentrer en jeu dans cette étude des considérations sur la nature du style employé (direct ou indirect), sur le langage dit affectif qui met en jeu le souhait, l'interrogation, la négation, la dénégation et d'autres notions.

À partir de deux textes journalistiques, nous avons tenté de montrer quelques pistes de recherche sur la caractérisation des textes pour l'extraction d'information, en particulier l'argumentation à partir des *tokens*, et dessiné quelques orientations de notre programme de recherche. Pour un examen de l'extraction de l'information en général, il va sans dire que l'extraction de la racine est primordiale et fera l'objet d'une étude à part.

14. Notamment avec un autre texte sur la même question, issu du même journal qui constitue une sorte d'intermédiaire entre celui que nous venons d'examiner et une position moins argumentative si l'on peut dire, et plus proche d'un rapport. À l'opposé, un texte sur le football qui a des traits argumentatifs serait à rapprocher de cet ensemble.

AnIsl 44 (2010), p. 39-52 Claude Audebert
Quelques réflexions sur la fréquence et la distribution des mots-outils ou "tokens" dans les textes arabes en vue de leur caractérisation dans le cadre de l'extraction d'information.