Des paires de lettres porteuses de sens dans le lexique coranique

Une méthode de lecture par unités bilatérales, et un premier test de contrôle indépendant des lectures proposées

Bechini Rami · chercheur indépendant · novembre 2026

Version de travail destinée à la relecture par des spécialistes. Code, données et graines aléatoires fournis pour reproduire chaque chiffre.

Résumé. La méthode décompose chaque mot du *rasm* coranique en paires de lettres : deux lettres voisines (consécutive), deux lettres séparées par une seule (noyau), ou par deux (super-noyau). Elle fait l'hypothèse que ces paires portent une part du sens. Cet article ne teste pas les lectures de versets qui en découlent, mais leur prémisse lexicale, dans une source écrite sans cette hypothèse : le dictionnaire de Lane (1863-1893). Sur 1279 racines trilitères du Coran (817281 couples), deux racines qui partagent une même paire ordonnée de lettres ont des définitions **1,21 fois** plus proches que le hasard quand la paire est consécutive dans les deux (z = 13,0), et **1,19 fois** quand elle forme un noyau dans les deux (z = 7,2). Ce chiffre tient compte de l'ordre alphabétique du dictionnaire. Deux lettres communes dans l'ordre inverse ne donnent que ×1,04, une seule lettre commune ×1,02. Sur neuf variantes de la mesure (trois sources, deux mesures, trois contrôles), l'effet des consécutives se retrouve partout (de ×1,18 à ×1,73), celui des noyaux huit fois sur neuf (de ×1,13 à ×1,47 ; la neuvième variante donne ×1,08, non significatif). Il est net en moyenne mais modeste : la paire explique une part du sens, pas tout le sens. Unité par unité, la mesure manque de puissance : les 238 unités nommées par la méthode ne se distinguent pas des autres paires. Un protocole à l'aveugle, prêt à l'emploi, est proposé pour tester la spécificité des notions attribuées aux unités.

Mots-clés : lexique arabe, racines bilitères, théorie des matrices et étymons, rasm, Coran, sémantique distributionnelle, test de permutation.

1. La question

L'idée que le lexique arabe repose, sous les racines de trois consonnes, sur des noyaux de deux consonnes n'est pas neuve. Ibn Jinnī (m. 1002) a exploré dans *al-Khaṣāʾiṣ* ce qu'il appelle le « grand dérivé » (*al-ishtiqāq al-akbar*) : des racines faites des mêmes consonnes, dans un ordre différent, partageraient un sens commun. Son contemporain Ibn Fāris (m. 1004), dans *Maqāyīs al-lugha*, ramène chaque racine à un ou quelques sens premiers (*aṣl*), et décrit une partie des racines de plus de trois lettres comme des racines composées. Au XX^{e} siècle, Georges Bohas a construit une théorie complète de cette intuition : les racines trilitères s'organisent autour d'« étymons » de deux consonnes, eux-mêmes rattachés à des « matrices » de traits phonétiques (Bohas 1997 ; Bohas et Dat 2007). Du côté quantitatif, Greenberg (1950) a montré que les consonnes d'une racine sémitique ne se combinent pas au hasard, et les travaux sur les « phonesthèmes » (Bergen 2004) ont montré qu'une suite de sons peut être associée à un champ de sens sans être un morphème.

Je reprends cette lignée sur trois points. D'abord, j'applique la décomposition au *rasm* (le tracé consonantique, sans voyelles ni points diacritiques ajoutés) de chaque mot du Coran, et non aux racines du dictionnaire. Ensuite, j'étends le principe aux lettres non voisines : deux lettres séparées par une lettre

forment un « noyau ». Enfin, j'attribue à chaque paire une notion, que je confronte à toutes ses occurrences dans le corpus. Cette méthode sert de base à une série de livres et de films (*NABA*).

Une telle méthode peut produire des lectures cohérentes sans que sa prémisse soit vraie : quand on lit un texte avec une notion assez souple, tout peut sembler coller. La question de cet article est donc plus étroite et plus dure : **indépendamment de toute lecture de verset, les paires de lettres sont-elles associées à du sens, dans une source qui n'a jamais entendu parler de la méthode ?** Si la réponse était non, les lectures n'auraient pas de fondement lexical. Si elle est oui, la prémisse est au moins plausible, et la discussion peut porter sur les notions elles-mêmes.

2. La méthode

2.1 Le corpus et le tracé

Le texte est celui de la lecture de Ḥafṣ d'après ʿĀṣim (6 234 versets). Chaque mot est ramené à son tracé : voyelles, signes de lecture et *shadda* sont retirés ; les formes de l'*alif* (ا أ إ آ ٱ) sont unifiées, ى est lu ي. On décompose ce tracé, jamais la racine du dictionnaire.

2.2 Trois niveaux, une seule loi

Une **consécutive** est une paire de lettres voisines (positions *i*, *i*+1). Un **noyau** naît de la fusion de deux consécutives qui partagent une lettre : la lettre partagée disparaît, il reste les positions *i* et *i*+2. Un **super-noyau** naît de même de deux noyaux (*i*, *i*+3). La loi ne dépend pas de la longueur du mot. Exemple : قَتَلَ se décompose en قت (consécutive), تل (consécutive) et قل (noyau). Dans يُقَٰتِلُونَكُمْ, les mêmes unités réapparaissent, avec d'autres qui viennent des lettres ajoutées.

2.3 Le glossaire

238 paires ont reçu un nom et une notion. Chaque notion a été établie en examinant les mots du corpus qui portent la paire, puis confrontée à toutes leurs occurrences. Exemples : سج *Confinement*, رض *Mœurs*, لح *Instance*, قت *Implémenter*, تل *Minutie*, قل *Intention*. Le glossaire complet et l'outil de décomposition sont publics (site ai-bed-coran.com ; script `decomposer.py`).

2.4 Ce qui peut être testé

La méthode repose sur trois affirmations de force croissante :

1. **P1 — la prémisse lexicale.** Une paire de lettres est associée à une part de sens partagée par les mots qui la portent.
2. **P2 — les notions.** La notion attribuée à chaque paire capte ce sens partagé, et non un sens assez vague pour aller à tout.
3. **P3 — les lectures.** Lire un verset avec les notions de ses unités donne un texte cohérent.

P1 se teste par le calcul, sans rien emprunter à la méthode (sections 3 et 4). P2 demande des juges humains qui ignorent la méthode (section 6). P3 dépend de P1 et P2 ; la section 5 en donne un exemple, sans le compter comme une preuve.

3. Test A : deux racines qui partagent une paire ont-elles des sens plus proches ?

3.1 Données

Les racines sont les racines trilitères du Coran pour lesquelles l'édition numérique du dictionnaire de Lane (Lane 1863-1893 ; corpus de Dukes et Habash 2010) donne un article : 1 279 racines. Les hamzas sont ramenées à ا, comme dans le tracé. Dans chaque article, je retire toute écriture arabe (pour qu'aucune lettre ne passe dans la mesure), puis le vocabulaire qui décrit la rédaction de Lane plutôt que le sens de la racine : termes de grammaire (*plural, measure, aorist, quiescent*…), renvois éditoriaux (*preceding paragraph*…) et sigles de ses sources. Sans ce nettoyage, deux racines de même forme grammaticale, par exemple à dernière lettre faible, paraîtraient proches parce que Lane les commente avec les mêmes mots. Les mots présents dans plus de 20 % des articles, ou dans un seul, sont écartés (vocabulaire final : 6 057 mots).

Deux sources servent de réplication : les résumés anglais des articles de Lane (un paragraphe par racine), et les gloses françaises d'une traduction mot à mot du Coran, réunies par racine (1 351 racines). Aucune de ces trois sources n'a été écrite avec les unités bilatérales en tête.

3.2 Mesure

Chaque racine devient un vecteur de mots pondérés (TF-IDF : fréquence sous-linéaire, pondération par la rareté ; Salton et Buckley 1988). La proximité de deux racines est le cosinus de leurs vecteurs : 0 quand leurs définitions n'ont aucun mot en commun, 1 quand elles sont identiques. Une seconde mesure projette les vecteurs sur 100 dimensions latentes (Deerwester et al. 1990), ce qui rapproche des synonymes qui ne partagent pas le même mot.

3.3 Catégories de couples

Chaque couple de racines est rangé dans une seule catégorie, de la plus forte à la plus faible : la même paire ordonnée en consécutive dans les deux (CC), la même paire en noyau dans les deux (NN), consécutive dans l'une et noyau dans l'autre (CN), deux lettres communes mais jamais dans le même ordre (REV), une seule lettre commune (ONE), aucune (ZERO). Les catégories REV et ONE sont les témoins décisifs : si le seul fait de partager des lettres rapprochait les sens (par la sonorité, par exemple), REV devrait valoir CC.

3.4 Hypothèse nulle

Si le sens n'avait rien à voir avec les lettres, on pourrait redistribuer les définitions entre les racines sans rien changer. On le fait 2 000 fois, et l'on compare la proximité moyenne observée dans chaque catégorie à la distribution obtenue (test de permutation, dans l'esprit de Mantel 1967). Les couples ne sont pas indépendants (une racine entre dans beaucoup de couples) : la permutation respecte cette structure, ce qu'un test classique ne ferait pas.

Un piège propre aux dictionnaires impose un second contrôle. Lane range ses articles par ordre alphabétique de racine : les racines qui partagent leurs deux premières lettres sont des articles voisins, rédigés à la suite, dans le même volume, parfois par le même continuateur. Leur style seul pourrait les rapprocher. La permutation **stratifiée** ne redistribue les définitions qu'entre racines de même première lettre. C'est le chiffre retenu dans le texte.

3.5 Résultats

Catégorie de couples	Exemple	Couples	Cosinus moyen	Rapport au hasard (global)	Rapport au hasard (même 1re lettre)	z (même 1re lettre)	p
CC · même paire, consécutive dans les deux	سفح / سفك	8 801	0,0220	×1,24	×1,21	13,0	< 0,001
NN · même paire, en noyau (lettres 1 et 3) dans les deux	خشع / خضع	2 393	0,0247	×1,39	×1,19	7,2	< 0,001
CN · même paire, consécutive dans l'une, noyau dans l'autre	قتل / عقل	8 322	0,0198	×1,11	×1,05	3,0	0,001
REV · deux lettres communes, jamais dans le même ordre	قتل / لقي	15 974	0,0182	×1,03	×1,04	3,0	0,001
ONE · une seule lettre commune	قتل / قرب	276 568	0,0183	×1,03	×1,02	5,5	< 0,001
ZERO · aucune lettre commune	قتل / سجد	505 223	0,0174	×0,98	×0,98	-7,3	—

Tableau 1. Proximité des définitions de Lane selon les lettres partagées (TF-IDF, 1 279 racines, 817 281 couples, 2 000 permutations). « Rapport au hasard » : cosinus moyen observé divisé par le cosinus moyen des permutations. p unilatéral ; « < 0,001 » : aucune des 2 000 permutations n'atteint la valeur observée.

Deux racines qui partagent une paire consécutive ont des définitions 1,21 fois plus proches que le hasard, ordre alphabétique neutralisé (z = 13,0). Le noyau fait autant : ×1,19 (z = 7,2). Les témoins restent presque plats : deux lettres en ordre inverse ×1,04, une lettre ×1,02. **Ce n'est donc pas le partage de lettres qui rapproche les sens, c'est le partage d'une paire ordonnée.**

Position de la paire consécutive	Couples	Rapport (global)	Rapport (même 1re lettre)	z (même 1re lettre)
lettres 1-2 dans les deux (articles voisins chez Lane)	2 111	×1,40	×1,23	8,3
lettres 2-3 dans les deux	2 791	×1,34	×1,32	12,4
lettres 1-2 dans l'une, 2-3 dans l'autre	3 899	×1,08	×1,10	4,5

Tableau 2. La paire consécutive selon sa position dans la racine.

Le tableau 2 confirme que le contrôle alphabétique était nécessaire : pour les paires en tête de racine (articles voisins chez Lane), l'effet passe de ×1,40 à ×1,23 une fois l'ordre neutralisé. Mais les paires en fin de racine, que l'ordre alphabétique disperse dans tout le dictionnaire, donnent ×1,32, contrôle ou non : l'effet ne vient pas de la rédaction. Une paire placée à des positions différentes dans les deux racines rapproche moins (×1,10). La méthode traite les unités sans égard à leur position ; ce résultat suggère que la position compte aussi, et mérite d'être intégré.

Catégorie	Part des couples parmi les 5 % les plus proches du dictionnaire	Rapport à 5 %
CC · même paire, consécutive dans les deux	7,7 %	×1,54
NN · même paire, en noyau (lettres 1 et 3) dans les deux	9,4 %	×1,89
CN · même paire, consécutive dans l'une, noyau dans l'autre	6,3 %	×1,26
REV · deux lettres communes, jamais dans le même ordre	5,5 %	×1,10
ONE · une seule lettre commune	5,3 %	×1,06
ZERO · aucune lettre commune	4,7 %	×0,95

Tableau 3. Une lecture plus parlante du même effet : la part des couples qui tombent parmi les 5 % de couples les plus proches de tout le dictionnaire (cosinus ≥ 0,059). Au hasard, on attendrait 5 %.

Couple	Type	Cosinus	Mots anglais communs les plus lourds (Lane)
سفح / سفك	CC	0,460	*pour, tear, blood, shed, forth*
سكت / سكن	CC	0,393	*still, quiet, allay, appeas, assuag*
فرح / مرح	CC	0,358	*exult, glad, joyful, rejoic, ungrateful*
رجس / نجس	CC	0,293	*unclean, filthy, dirty, perhap*
بذر / ذرو	CC	0,287	*scatter, sow, grain, dispers, seed*
عوذ / لوذ	CC	0,253	*preservation, refuge, protection, sought, recourse*
ثقب / نقب	CC	0,242	*hol, perforat, pierc, hole, bor*
خشع / خضع	NN	0,382	*stoop, sett, voice, adn, eyn*
ابد / امد	NN	0,366	*duration, limit, existence, everlast, space*
لمح / لوح	NN	0,330	*shone, glisten, gleam, slight, look*
شهب / شيب	NN	0,284	*snow, white, frost, hoar, cold*
بغت / بهت	NN	0,236	*unawar, unexpect, sudden, surprise, cause*
قتر / قدر	NN	0,222	*subsistence, scant, expenditure, parsimoniou, straiten*
سبت / سكت	NN	0,158	*silent, quiet, ceas, still, motionles*

Tableau 4. Exemples, choisis parmi les 25 couples les plus proches de chaque catégorie, avec les mots anglais de Lane qui portent leur proximité. La racinisation tronque les mots (*appeas* pour *appease*).

3.6 Robustesse

Source et mesure	Racines	CC	NN	CN	REV	ONE
Lane, articles · TF-IDF (mesure principale)	1279	×1,21 (z 13,0)	×1,19 (z 7,2)	×1,05 (z 3,0)	×1,04 (z 3,0)	×1,02 (z 5,5)
Lane, articles · analyse latente (100 dim.)	1279	×1,18 (z 11,6)	×1,13 (z 5,3)	×1,05 (z 3,0)	×1,04 (z 2,9)	×1,02 (z 5,6)
Lane, résumés · TF-IDF	1584	×1,52 (z 14,3)	×1,37 (z 5,4)	×1,12 (z 3,1)	×1,04 (z 1,4)	×1,02 (z 3,1)
Lane, résumés · analyse latente	1584	×1,29 (z 9,5)	×1,21 (z 3,8)	×1,11 (z 3,3)	×1,02 (z 1,0)	×1,02 (z 3,2)
Gloses françaises du corpus · TF-IDF	1351	×1,73 (z 7,7)	×1,47 (z 2,8)	×1,17 (z 1,7)	×1,18 (z 2,2)	×1,01 (z 0,7)
Gloses françaises du corpus · analyse latente	1351	×1,36 (z 4,0)	×1,08 (z 0,5)	×1,19 (z 2,0)	×1,15 (z 2,0)	×1,01 (z 0,8)
Lane · sans racines à lettre faible (ا و ي)	861	×1,20 (z 10,1)	×1,19 (z 5,5)	×1,05 (z 2,3)	×1,05 (z 3,0)	×1,02 (z 3,6)
Lane · sans racines à lettre redoublée	1272	×1,21 (z 13,3)	×1,19 (z 6,8)	×1,05 (z 3,0)	×1,04 (z 2,9)	×1,02 (z 5,4)
Lane · mots pris tels quels (sans racinisation)	1279	×1,22 (z 13,5)	×1,20 (z 7,0)	×1,05 (z 3,4)	×1,04 (z 3,2)	×1,02 (z 5,5)

Tableau 5. Rapport au hasard (permutation stratifiée par première lettre) et score z, pour chaque source et chaque variante.

L'effet des consécutives (CC) ne dépend ni de la source, ni de la mesure, ni des racines à lettre faible ou redoublée, ni de la racinisation : il est significatif dans les neuf variantes. Celui des noyaux (NN) l'est dans huit ; la seule exception est celle des gloses françaises mesurées par analyse latente, où il tombe à ×1,08 (z = 0,5). Une différence apparaît aussi entre les sources : dans les gloses françaises, deux lettres en ordre inverse rapprochent un peu les sens (REV ×1,15 à ×1,18), ce que les deux sources tirées de Lane ne montrent pas. Ce résultat va dans le sens du « grand dérivé » d'Ibn Jinnī, sans changer la hiérarchie : partout, la paire dans le même ordre rapproche davantage que la paire inversée.

Ce que dit le test A. Dans le lexique coranique, une paire ordonnée de lettres, consécutive ou en noyau, est associée à du sens partagé (consécutive : neuf variantes sur neuf ; noyau : huit sur neuf). L'association est très improbable par hasard. Elle est aussi modeste en moyenne : beaucoup de couples qui partagent une paire n'ont pas de définitions visiblement proches. Le test établit P1 au niveau du lexique entier ; il ne dit rien, à lui seul, de chaque unité ni de chaque notion.

4. Test B : unité par unité

Pour chaque paire ordonnée de lettres, je réunis les racines qui la portent et je mesure la proximité moyenne de leurs définitions, comparée à celle de groupes de même taille tirés au hasard (5 000 tirages). Les seuils tiennent compte du grand nombre de paires testées (correction de Benjamini et Hochberg 1995).

Paires portées en…	Paires testées (≥ 4 racines)	Au-dessus du hasard	Signe (p)	Significatives (BH, 5 %)	z médian nommées / non nommées	Nommées > non nommées (Mann-Whitney)
consécutive	286	175 (61 %)	0,00009	7	0,50 / 0,40	0,50 (p = 0,54)
noyau	132	86 (65 %)	0,00032	11	0,59 / 0,65	0,50 (p = 0,53)
consécutive ou noyau	373	246 (66 %)	< 0,00001	9	0,53 / 0,40	0,52 (p = 0,28)

Tableau 6. Cohérence de chaque paire. « Signe » : probabilité d'avoir autant de paires au-dessus du hasard si chacune avait une chance sur deux. « Nommées » : les 238 unités du glossaire ; « non nommées » : les autres paires.

Deux paires sur trois sont plus cohérentes que le hasard (246 sur 373), ce qui serait très improbable si les paires ne portaient rien. Mais peu d'entre elles le sont assez, isolément, pour passer le seuil : un groupe de quatre à vingt racines laisse peu de puissance à une mesure aussi grossière. Surtout, **les 238 unités nommées par la méthode ne sont pas plus cohérentes que les paires qu'elle n'a pas nommées** (probabilité de supériorité 0,52 ; p = 0,28). Ce résultat est cohérent avec l'idée que toutes les paires portent du sens, pas seulement celles qui ont déjà reçu un nom ; il signifie aussi que ce test ne valide pas le choix des 238.

Unité	N°	Notion (glossaire)	Racines	z	p	Racines porteuses (Lane)
لت	82	Anarchisme	4	8,33	0,0002	الت لفت لوت ليت
اد	214	Restitution	6	4,49	0,0008	ابد احد ادم امد اودايد
رض	230	Mœurs	11	4,39	0,0006	ارض حرض رضع رضو ركض رمض روض عرض فرض
عد	46	Aptitudes	17	4,32	0,0002	بعد رعد سعد صعد عبد عتد عدس عدل عدو
عو	149	Transducteur	21	4,14	0,0002	دعو عتو عثو عدو عرو عزو عشو عصو عضو
رع	61	Bouleversement	19	3,94	0,0002	جرع ذرع ربع رتع رجع رضع رعب رعد رفع
دو	87	Registre	13	3,67	0,0012	بدو دحو دعو دلو دمو دنو دور دول دوم
ال	155	Destination	14	3,37	0,0018	ابل اثل اجل اصل افل اكل الت الف الم
حو	140	Contenu	14	3,15	0,0028	حفو حوب حوت حوج حوذ حور حوز حوش حوط
بي	125	Manifeste	6	3,02	0,0064	بيت بيد بيض بيع بين جبي
عل	122	Impartialité	20	2,96	0,0042	بعل جعل شعل عتل عجل عدل عزل عسل عضل
ير	198	Trajectoire	10	2,76	0,0090	حير خير سير صير ضير طير عير غير مير
Unités du chapitre 10 et autres unités citées						
سج	1	Confinement	6	0,65	0,2366	سجد سجر سجل سجن سجو سرج
ول	90	Déléguer	18	0,63	0,2549	اول بول حول خول دول زول سول طول عول
تل	117	Minutie	6	0,50	0,2823	بتل تلو رتل عتل فتل قتل
قت	49	Implémenter	6	0,18	0,3843	قتر قتل قنت قوت مقت وقت
قل	30	Intention	12	-1,28	0,9154	بقل ثقل عقل قبل قتل قفل قلب قلد قلع

Tableau 7. Les unités nommées les plus cohérentes dans le dictionnaire de Lane, puis les unités citées dans cet article. Prudence pour les unités qui contiennent une lettre faible (ا, و, ي) : leurs racines partagent souvent une même forme grammaticale, et une part de leur cohérence peut venir de là malgré le nettoyage du vocabulaire de grammaire.

L'unité رض (*Mœurs*) se distingue (z = 4,39) : ارض حرض رضع رضو ركض رمض روض عرض. En noyau, لح (*Instance*) réunit لفح لقح لمح لوح, que Lane définit tous par l'éclat, la lueur, le regard bref ou le souffle brûlant (z = 8,60). À l'inverse, les unités قت, تل et قل ne se distinguent pas du hasard avec cette mesure (z = 0,18, 0,50 et -1,28). J'y reviens à la section 5 : la mesure compare les mots concrets des définitions, et ne peut pas voir une notion abstraite que des racines partagent sans partager un seul mot.

5. Étude de cas : قتل, les 122 versets

La racine قتل apparaît 170 fois, dans 122 versets. Le sens courant du dictionnaire est « tuer ». La méthode la lit par ses unités : قت *Implémenter* et تل *Minutie* en consécutives, قل *Intention* en noyau. D'où la notion : *mettre en œuvre avec minutie, selon une intention*.

Le dictionnaire lui-même ouvre cette piste. Dans la partie de l'article de Lane que contient l'édition numérique, le premier sens donné est figuré : pour قَتَلَ الشَّيْءَ, « *he knew it* [...] *completely,* [...] *or thoroughly, very well, or superlatively well* » (il la connut à fond), avec l'expression قَتَلَهُ عِلْمًا, sur l'autorité du *Qāmūs*, du *Tāj al-ʿArūs* et d'al-Bayḍāwī. La référence à al-Bayḍāwī porte sur un verset du Coran, que Lane cite dans la numérotation de Flügel (iv. 156) : c'est 4:157 dans celle de Ḥafṣ, وَمَا قَتَلُوهُ يَقِينًا.

L'exégète y lit donc déjà قَتَلُوهُ comme une connaissance menée à son terme. Le dictionnaire connaît un emploi de قتل qui désigne la maîtrise d'une chose, et non sa suppression.

Les 122 versets ont été relus un par un avec cette notion. Selon ma lecture, 50 la reçoivent directement, et 72 au terme d'un raisonnement qui s'appuie sur le contexte (répertoire complet en annexe du chapitre 10). **Ce décompte n'est pas une mesure indépendante** : c'est le mien, et c'est moi qui ai proposé la notion. Il montre que la notion ne se heurte à aucun verset ; il ne montre pas qu'une autre notion ferait moins bien. C'est précisément ce que le protocole de la section 6 doit établir.

Le test B n'apporte pas d'appui à قت, تل et قل : les racines qui les portent (par exemple قتر قتل مقت وقت : *être avare*, *tuer*, *détester*, *le temps*) n'ont aucun mot en commun chez Lane. Ce n'est pas une réfutation, puisqu'une notion abstraite (mesurer, mettre en œuvre, compter juste) peut relier des sens concrets sans mot commun. Mais tant qu'une mesure plus fine ou des juges humains ne l'ont pas montré, l'appui statistique de ces trois unités reste à établir.

6. Test C : un protocole à l'aveugle pour les notions

Pour P2, il faut des juges qui ignorent la méthode. Chaque question montre les sens de dictionnaire de cinq racines qui portent la même paire, sans les lettres ni l'arabe. Le juge choisit, parmi quatre notions, celle qui réunit le mieux ces sens. Une seule est la notion du glossaire ; les trois autres sont des notions d'autres unités, tirées au hasard. Si les notions étaient vagues, les juges répondraient au hasard : une fois sur quatre.

Le questionnaire est prêt : 30 questions sur des unités tirées au sort, et 6 questions témoins faites de cinq racines sans paire commune, qui mesurent le biais des juges. Deux précautions : aucune racine dont la définition contient le mot de la notion n'est montrée (la réponse serait donnée), et la clé est conservée à part jusqu'à la correction.

Plan d'analyse, fixé avant toute réponse. Critère : la proportion de bonnes réponses aux questions d'unités, tous juges réunis, comparée à 25 % (loi binomiale, unilatéral, seuil 5 %). Avec 10 juges (300 réponses), il faudra au moins 88 bonnes réponses (29,3 %). Si les notions donnent en réalité 35 % de bonnes réponses, ce plan le détectera dans 98 % des cas ; à 30 %, dans 62 % des cas. Les questions témoins doivent rester proches de 25 %.

7. Discussion

7.1 Ce qui est établi

Dans le lexique du Coran, deux racines qui partagent une paire ordonnée de lettres ont des sens plus proches que le hasard, dans trois sources indépendantes de la méthode, avec un contrôle de l'ordre du dictionnaire. L'ordre des lettres compte : deux lettres en ordre inverse rapprochent beaucoup moins, ou pas du tout selon la source. Le résultat le plus propre à cette méthode est celui du noyau : deux racines qui partagent leur première et leur dernière lettre se rapprochent autant que deux racines qui partagent deux lettres voisines. La lettre du milieu peut changer, la paire reste porteuse.

7.2 Ce qui ne l'est pas encore

- **Le tracé entier.** Le test porte sur les racines. La méthode décompose aussi les lettres de dérivation et de flexion (préfixes, suffixes, article) ; que ces paires-là portent du sens n'est pas testé ici.

- **Les notions.** Le test A ne dit pas que la notion de chaque unité est la bonne. Le test B ne distingue pas les unités nommées des autres. C'est l'objet du protocole à l'aveugle.
- **Les lectures de versets.** Elles restent des propositions, dont la valeur dépend des deux points précédents.

7.3 Limites de la mesure

- L'édition numérique de Lane est inégale : certains articles sont complets, d'autres réduits à un fragment (celui de قتل, par exemple). Cette inégalité ajoute du bruit, ce qui joue contre l'hypothèse et non pour elle.
- TF-IDF ne voit que les mots partagés. Il ne voit ni les synonymes éloignés, ni les métaphores, ni les notions abstraites. Les effets mesurés sont donc vraisemblablement des minorants.
- Lane a lui-même lu les grammairiens arabes, dont certains rapprochaient déjà des racines voisines. Une partie de la proximité peut venir de cette tradition. La réplication sur les gloses d'une traduction mot à mot, établies verset par verset, limite ce biais sans l'exclure.
- Seules les racines attestées dans le Coran sont étudiées. L'extension au lexique arabe entier, puis aux autres langues sémitiques, est l'étape suivante.

7.4 Suites

Trois chantiers suivent : faire passer le protocole à l'aveugle ; refaire le test A avec une mesure sémantique plus fine et sur d'autres dictionnaires (*Lisān al-ʿArab*, *Maqāyīs*) ; confronter la position des paires, que le tableau 2 fait apparaître, au principe de la méthode. Je soumets ces résultats à la critique : toute objection précise sur la mesure ou sur les contrôles sera prise en compte dans la version suivante.

8. Reproduire

Tout le calcul tient en quelques scripts Python (bibliothèque NumPy), dans le dossier `etude-controle` : `commun.py` (données, nettoyage, vecteurs), `test_A_paires.py`, `test_A_complements.py`, `test_B_unites.py`, `test_C_questionnaire.py` et `test_C_correction.py`. Les graines aléatoires sont fixées (20261114, 20261115, 20261116) : chaque chiffre de cet article se retrouve à l'identique. Les données sont le texte de Ḥafṣ, l'édition numérique de Lane (Dukes et Habash 2010, licence GPL-3.0) et le glossaire des 238 unités.

Références

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289-300.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2), 290-311.

Bohas, G. (1997). *Matrices, étymons, racines. Éléments d'une théorie lexicologique du vocabulaire arabe*. Louvain et Paris : Peeters.

Bohas, G. et Dat, M. (2007). *Une théorie de l'organisation du lexique des langues sémitiques : matrices et étymons*. Lyon : ENS Éditions.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Dukes, K. et Habash, N. (2010). Morphological annotation of Quranic Arabic. *Proceedings of LREC 2010*.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, 6(2), 162-181.

Ibn Fāris, Aḥmad. *Maqāyīs al-lugha*, éd. ʿAbd al-Salām M. Hārūn. Le Caire.

Ibn Jinnī, Abū l-Fatḥ ʿUthmān. *Al-Khaṣāʾiṣ*, éd. Muḥammad ʿAlī al-Najjār. Le Caire.

Lane, E. W. (1863-1893). *An Arabic-English Lexicon*. Londres : Williams and Norgate.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2), 209-220.

Salton, G. et Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.