Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Lost In Variation : extraction non-supervisée de motifs lexico-syntaxiques dans des textes en moyen arabe

Julien Bezançon[1, 5] Rimane Karam[1, 2, 3, 4] Gaël Lejeune[1, 5]
(1) CERES, (2) Orient & Méditerranée, (3) LiPoL, (4) Ifpo, (5) STIH
Sorbonne Université, 28 rue Serpente 75006 Paris, firstname.lastname@sorbonne-universite.fr

RÉSUMÉ

Contrairement à l'arabe standard moderne ou à certains dialectes de l'arabe, le moyen arabe a peu été étudié en TAL. Pourtant, cette famille de variétés présente un défi majeur : elle mêle des traits de standard et des traits de dialecte en plus de posséder des caractéristiques qui lui sont propres. Nous présentons ici une méthode pour identifier, extraire et classer les variantes de 13 formules du moyen arabe, relevées manuellement. Ces formules proviennent des neuf premiers tomes du corpus SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ, un corpus de littérature populaire rédigé dans une variété de moyen arabe proche du dialecte damascène. Nous classons 20 386 séquences en se fondant sur leur similarité à plusieurs niveaux avec les formules étudiées. Ce classement nous permet d'observer que ces formules varient sur les plans lexical, morphologique et graphique tout en restant strictement invariables sémantiquement et syntaxiquement.

ABSTRACT

Lost in Variation : Unsupervised Mining of Lexico-syntactic Patterns in Middle Arabic Texts

Modern Standard Arabic and some dialects of Arabic have been extensively studied in NLP, Middle Arabic to the contrary has not attracted much attention. However, it offers interesting challenges for NLP since it is characterized by variation and mixes standard features, colloquial ones, as well as features of its own. Here, we introduce a methodology to identify, extract and classify variations of 13 selected formulas. These formulas come from the nine first booklets of SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ, a corpus of Damascene popular literature written in Middle Arabic. We classified 20,386 sequences according to their similarity with the original formulas on multiple linguistic layers. We noticed that the variations in these formulas occur in a lexical, morphological and graphical level, but in opposition, the semantic and syntactic levels remain strictly invariable.

MOTS-CLÉS : fouille de textes, variation, moyen arabe, similarité, alignement de séquences.

KEYWORDS: text mining, linguistic variation, middle arabic, similarity, sequence alignment.

ARTICLE : **Accepté à** WACL (*Workshop* de la conférence COLING, 2025).
Lien : https://aclanthology.org/2025.wacl-1.3/.

1 Introduction

Trois variétés principales d'arabe ont été étudiées en TAL : l'arabe classique, l'arabe standard moderne (MSA) et certains dialectes. L'arabe classique a fait l'objet de quelques travaux, mais les recherches en TAL se sont concentrées sur le MSA et les dialectes. À notre connaissance, le moyen arabe n'a

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : traductions d'articles publiés, pages 36–49.

jamais été étudié dans une perspective de TAL. Pourtant, aborder le moyen arabe par le prisme du TAL est intéressant à bien des égards. Le moyen arabe est « [...] un ensemble de registres linguistiques, qui s'inscrivent dans le *continuum* linguistique arabe, borné par deux pôles de statut différent, la langue dialectale d'une part, la langue standard d'autre part, et qui se caractérisent par le fait qu'ils ne sont ni dialectaux, ni standard » (Lentin, 2003). Cette famille de registres est marquée par la variation linguistique (Zack & Schippers, 2012). Étudier le moyen arabe en TAL permet de traiter conjointement plusieurs variétés, ce qui est utile pour le traitement d'autres états de langue puisque les textes écrits en arabe sont rarement rédigés dans une seule variété (Katz & Diab, 2011).

La nature mixte du moyen arabe et sa propension à la variation offrent des défis originaux pour la linguistique et le TAL (Section 2). Ainsi, des formules comme « فز واثب على الاقدام » (il se redressa en bondissant sur ses pieds) peuvent aussi s'écrire « نهض واثب على الاقدام » (il se leva en bondissant sur ses pieds) ou avec la variation graphique « فذ واثب على الاقدام » (où *fzz* est remplacé par *fḏḏ*). À cela s'ajoutent d'autres enjeux de TAL arabe plus généraux : l'ambiguïté orthographique, la richesse morphologique et le bruit (Habash, 2010).

Nous proposons ici une méthode pour étudier des corpus comprenant plusieurs variétés de langue, avec l'objectif d'identifier les variations d'une même formule. Nous travaillons sur le corpus SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ, composé de 53 843 phrases. Nous y recherchons les variantes de 13 formules relevées par une linguiste afin d'analyser leurs variantes. Pour ce faire, nous utilisons de l'alignement au grain token pour repérer des motifs lexico-syntaxiques et des mesures de similarité pour détecter les variations de nos formules. Notre travail se rapproche de l'identification d'expressions multi-mots (EMM) telle que définie dans Constant *et al.* (2017) en ce sens que les EMM sont des séquences conventionnelles et idiomatiques (Sag *et al.*, 2002). Nous approfondissons la méthodologie de Bezançon & Lejeune (2023) pour identifier et extraire des EMM et des EMM défigées. Nous présentons le moyen arabe dans la Section 2 puis le corpus textuel et les formules analysées dans la Section 3. Ces formules sont des séquences de mots fréquentes qui apparaissent régulièrement dans le corpus. Nous présentons notre méthode d'identification et d'extraction des variantes de ces formules dans la Section 4. Enfin, nous analysons des variantes observées dans la Section 5.

2 Le moyen arabe : une langue non standardisée

La langue arabe est souvent divisée en deux variétés : le standard et les dialectes. Cette situation linguistique, appelée diglossie, a été théorisée par Charles Ferguson : la variété dite « haute » fait référence au standard, par opposition aux dialectes, la variété « basse » (Ferguson, 1959). En réalité, cette catégorisation n'est pas aussi binaire : Ferguson avait déjà reconnu l'existence de variétés intermédiaires. Des recherches ultérieures définissent ces variétés situées entre les deux pôles de la diglossie sous le terme de moyen arabe (Blau, 1982). Cette famille de variétés regroupe l'ensemble des registres intermédiaires mêlant des traits standards, des traits dialectaux ainsi que des traits qui leur sont propres – ni standards ni dialectaux – et qui appartiennent à un troisième pôle (Larcher, 2001).

Tout un pan de la littérature arabe a été rédigé en moyen arabe et *a fortiori* la littérature populaire, comme c'est le cas pour les MILLE ET UNE NUITS (Lentin, 2012). La recension damascène de SĪRAT BAYBARṢ, sur laquelle nous travaillons, en est un autre exemple. Elle est rédigée dans un état de langue très proche des dialectes levantins, mais présente des traits propres au standard ainsi que des traits qui ne relèvent d'aucun des deux pôles. Par exemple, le relatif « *alladī* » au masculin singulier demeure invariable quels que soient le genre et le nombre de son antécédent (Lentin, 2012).

Le caractère mixte et hybride du moyen arabe rend difficile l'utilisation d'outils tels que les étiqueteurs morphosyntaxiques, généralement conçus pour une variante particulière (standard ou dialecte). Il existe en effet une grande variété d'outils pour le MSA, tels que des outils de segmentation (Abdelali *et al.*, 2016) ou des étiqueteurs morphosyntaxiques (Zalmout *et al.*, 2018; Pasha *et al.*, 2014). Il existe aussi des outils spécifiques à des dialectes, tels que l'égyptien (Zalmout *et al.*, 2018; Samih *et al.*, 2017) ou les dialectes du golfe (Alharbi *et al.*, 2018; Khalifa *et al.*, 2017). Nous recensons également des outils pouvant traiter simultanément plusieurs dialectes (Darwish *et al.*, 2018; Al-Shargi *et al.*, 2016). En revanche, il n'existe pas d'outils dédiés au moyen arabe. Nous utiliserons ici CAMELTOOLS (Obeid *et al.*, 2022), un outil multi-dialectal qui couvre le MSA, l'égyptien, les dialectes du golfe et le levantin, pour étiqueter et lemmatiser notre corpus (Section 4). A défaut d'un traitement parfait du moyen arabe, nous pensons que cet outil devrait être capable d'identifier correctement ce qui se rapproche du MSA et des dialectes susmentionnés.

3 Description du jeu de données

3.1 Le corpus de textes

SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ est un cycle épique en prose d'époque ottomane. Il était avant tout destiné à la performance : les conteurs du Levant, les *hakawātī*-s, mémorisaient et récitaient les aventures de Baybars dans des cafés ou des demeures, par cœur ou à l'aide de cahiers. Dans cet article, nous utilisons la recension damascène de SĪRAT BAYBARṢ (Anonymous, 2000 2022). Ce corpus composite regroupe trois séries de cahiers de conteurs, mis par écrit par plusieurs scribes entre les XVIII^{e} et XX^{e} siècles puis rassemblés par des conteurs de Damas [1]. Nous centrons notre étude sur les 90 premiers cahiers du manuscrit Abu Ahmad, du conteur éponyme. Il est constitué de 183 cahiers, mais seuls les 90 premiers ont été transcrits numériquement. Dans la version critique éditée (Anonymous, 2000 2022), ils ont été divisés en 9 volumes de 10 cahiers chacun. Le Tableau 1 montre plusieurs statistiques sur chacun des volumes. Nous pouvons remarquer que le *Type-to-Token Ratio* (TTR) est très bas pour l'ensemble du corpus (7,06 %), ce qui peut indiquer qu'un grand nombre de constructions se répètent. Le tableau 2 montre une entrée traitée automatiquement (cf. Section 4.1).

Vol.	# Tokens	# Phrases	Moy.	TTR
1	94 315	5 679	16,61	16,41
2	100 408	6 482	15,49	15,00
3	118 986	6 093	19,53	15,23
4	92 744	4 389	21,13	16,20
5	105 081	5 562	18,89	15,15
6	106 817	6 515	16,40	14,46
7	119 921	7 504	15,98	15,09
8	82 691	5 235	15,80	17,12
9	107 972	6 384	16,91	14,89
All	928 935	53 843	17,25	07,06

TABLE 1 – Statistiques pour chaque volume du corpus.

sentence :	"فقال لى :والله ، انا احبك حباً شديد . "
id :	"27434"
tokens :	["فقال", "لى", " :", "والله", ",", "انا", "احبك", "حباً", "شديد", "."]
pos tags :	["verb", "prep", "punc", "noun_prop", "punc", "pron", "verb", "noun", "noun_prop", "punc"]
lemmas :	["قال", "لِ", " :", "اللَّه", ",", "أَنا", "أَحَبّ", "حُبّ", "شَدِيد", "."]

TABLE 2 – Exemple de phrase traitée linguistiquement.

La langue de SĪRAT BAYBARṢ est très proche du dialecte damascène, mais comprend aussi des marques d'autres dialectes. Certains personnages, par exemple, sont moqués sur leur manière de parler, dans un dialecte déformé ou caricatural. Ces deux strates de variation combinées – le moyen arabe damascène et les autres dialectes-idiolectes de la SĪRA – rendent difficiles l'analyse statistique de ce texte, *a fortiori* en l'absence d'outils spécifiquement dédiés au moyen arabe.

1. https://lipol.hypotheses.org/1310

3.2 Les formules

Ce sont des séquences de mots de la SĪRA qui apparaissent régulièrement et dans des contextes précis. Comme le montre J. P. Guillaume, chacune des occurrences d'une formule porte toujours la même signification, malgré les variations linguistiques (Guillaume, 2004). Si elles ne touchent pas la progression narrative, elles peuvent indiquer un changement soudain d'humeur chez un personnage, une séquence d'ouverture ou de clôture d'une situation (le jour qui se lève, la nuit qui tombe…). La régularité et le contexte d'apparition de ces séquences les rendent facilement repérables par le lecteur (ou l'auditeur) malgré les variations. Milman Parry définit « un groupe de mots régulièrement employé sous les mêmes conditions métriques pour exprimer une idée essentielle donnée »[2] sous le terme de formule (Parry, 1930). Son corpus de référence est la poésie homérique, un texte donc versifié. Bien qu'il nous soit parvenu sous une forme écrite, il est au départ ancré dans la tradition orale, tout comme notre corpus : l'oralité est en effet un élément important de la SĪRA, qui était initialement destinée à la performance (voir Section 3.1). Par ailleurs, bien que notre corpus soit principalement en prose, il est ponctué de séquences de mots dont l'usage est proche des formules d'Homère comme « Quand parut l'Aurore aux doigts de rose ». Dans la SĪRA, ces formules sont souvent utilisées dans un contexte de *sajᶜ* (prose rimée) : elles ne sont pas contraintes par des règles de versification, mais elles ne relèvent pas tout à fait de la prose, d'autant plus qu'elles ont tendance à provoquer une série d'autres formules en prose rimée à la suite.

D'autres contraintes que la versification, des contraintes linguistiques ou stylistiques, s'appliquent aux formules de SĪRAT BAYBARṢ. Deux éléments essentiels composent ces formules selon Parry : la régularité et l'expression d'une idée « sans réfléchir »[3] ce qui convient parfaitement aux formules de notre corpus. Elles représentent un point d'étape, pour le poète/scribe comme pour l'auditeur/le lecteur ; leur présence et leurs variations renseignent ainsi sur la langue du texte. Notre objectif est d'observer comment ces variations se forment à l'intérieur d'une formule, avec l'hypothèse qu'elles ne se forment pas aléatoirement, mais qu'elles suivent un schéma de variations plus général.

Treize formules ont été sélectionnées par une experte de SĪRAT BAYBARṢ. Pour des raisons d'intelligibilité, nous en présentons trois dans cet article, afin d'en donner une analyse détaillée :

1. غضب غضباً شديد
   (*ġḍb ġḍban šdīd*, « il se mit dans une grande colère »)
2. لما سمع فلان من فلان ذلك الكلام
   (*lmmā smᶜ flān mn flān ḏlk al-klām*, « lorsque A eut entendu de la part de B ces paroles »)
3. قلب الضيا بعينه ظلام
   (*qlb aḍ-ḍyā b-ᶜynh ẓlām*, « la lumière dans son oeil se transforma en ténèbres »)

La formule (1) indique une émotion très forte, la colère, qui résulte d'une situation ou du comportement d'un autre personnage. La formule (2) apparaît après qu'un personnage a tenu des propos qui en ont affecté un autre (positivement ou non). Souvent, elle est directement suivie par une réaction de ce personnage, ou par un changement soudain d'humeur. Quant à la formule (3), elle indique justement un changement d'humeur soudain et brutal, qui survient en réaction aux propos d'un personnage. Ces deux formules se retrouvent souvent l'une après l'autre. Nous souhaitons retrouver automatiquement, pour chaque formule, ses variantes, c'est-à-dire des séquences similaires mais non identiques : par exemple, pour 1, nous souhaiterions identifier (a.), (b.) et (c.) :

2. *"[...] a group of words which is regularly employed under the same metrical conditions to express a given essential idea"*
3. *"without second thought"*,(Parry, 1930)

a. غضبوا غضباً شديد
(*ġḍbū* ***ġḍban*** *šdīd*, « ils se mirent dans une grande en colère »)

b. غضبان غضبا شديد
(***ġḍbān*** *ġḍbā* *šdīd*, « il est dans une grande colère »)

c. وفرح فرحاً شديد
(***w-frḥ frḥan*** *šdīd*, « et il se mit dans une grande joie »)

Les variations peuvent être morphologiques avec ici un impact sur le verbe quand *ġḍbū* dans l'exemple (a.) remplace *ġḍb*. Elles peuvent aussi être d'ordre graphique comme c'est le cas pour l'exemple (b.) : la marque du *tanwīn* (nunation, marque de l'indéfini) est absente dans *ġḍbā* même si le *ʾalif* est indiqué, tandis que l'exemple (a.) le note dans *ġḍban*. Enfin, ces variations peuvent être lexicales, modifiant le lexème sans changer la structure de la séquence, comme le montre l'exemple (c.) où le verbe *ġḍb* employé dans les variantes (a.) et (b.) (se mettre en colère) devient *frḥ* (se mettre en joie).

4 Méthode de recherche de variations de formules

4.1 Traitement du moyen arabe

Nous utilisons CAMELTOOLS pour (i) tokeniser, (ii) étiqueter morpho-syntaxiquement et (iii) lemmatiser notre corpus[4]. Afin d'évaluer l'étiquetage, nous avons annoté manuellement 71 phrases (1 037 tokens) en limitant le nombre d'étiquettes à 5 (nom, préposition, nombre, ponctuation et verbe). La précision de CAMELTOOLS sur ces phrases est de 91,99 %, ce qui semble peu, étant donné le petit nombre d'étiquettes. Par exemple, le token « شديد » (*šdīd*) du Tableau 2 est étiqueté comme nom propre (*noun_prop*) alors qu'il s'agit d'un adjectif. Malgré ces observations, nous avançons que les annotations produites par CAMELTOOLS peuvent nous servir de base pour la suite de ce travail.

4.2 Association de séquences et classement des candidats

En premier lieu, nous comparons les phrases du corpus avec les 13 formules étudiées pour détecter des paires potentiellement intéressantes. Nous calculons la similarité cosinus (au token) de chaque paire possible phrase-formule et ne conservons que les paires dépassant un seuil de $0,1$. Par exemple, la phrase « غضب غضبا شديد » (il se mit dans une grande colère) et la formule « انا احبك حباً شديد » (je t'aime d'un grand amour) constituent une paire intéressante (similarité de 0,87). Nous approfondissons ensuite la méthode proposée par Bezançon & Lejeune (2023) pour détecter des EMM défigées en français afin de l'utiliser pour détecter nos variantes. Nous en décrivons ici les différentes étapes.

Alignement : Nous alignons chaque token de chaque phrase avec un token de la formule à laquelle elle a été associée, comme illustré ci-dessous. Les tokens non-alignés sont mis en évidence par ce procédé, comme c'est le cas pour le token « بعينه » , remplacé par le bigramme « فى وجهه »[5] :

Segmentation : Nous isolons la plus longue séquence de tokens commune de la paire phrase-formule qui partage le même premier et le même dernier token. Sur l'exemple donné ci-dessus, la phrase entière serait isolée puisqu'elle partage le même premier et dernier token avec la formule

4. Les scripts utilisés sont disponibles ici : https://github.com/JulienBez/ASMR, Le modèle levantin de CAMELTOOLS n'était pas disponible au moment de nos expériences

5. alignements effectués avec BIOPYTHONhttps://biopython.org/

قلب	الضيا	-	-	بعينه	ظلام
قلب	الضيا	فى	وجهه	-	ظلام

(respectivement « قلب » et « ظلام »). De plus, la phrase entière correspond à une variante de la formule renseignée. Nous notons enfin qu'une phrase peut avoir plus d'une séquence isolée pour une même formule.

Similarité : Pour chaque couche d'information linguistique (tokens, morphosyntaxe et lemmes), nous vectorisons chaque paire afin de calculer son score de similarité cosinus. Ce score est compris entre 0 et 1 : un score élevé indique que la séquence se rapproche de la formule.

Classement : Nous calculons la moyenne des similarités de chaque couche pour classer nos séquences. Le Tableau 3 présente la comparaison d'une formule avec une séquence pour chaque couche. En étudiant plusieurs couches d'information linguistique, nous espérons accroître l'efficacité de notre méthodologie.

Couche	Formule	Séquence	Score
Tok.	قلب الضيا بعينه ظلام	قلب الضيا فى وجهه ظلام	0.67
Lem.	قَلْب الضيا عَيْن ظَلام	قَلْب الضيا فِي وَجْه ظَلام	0.67
Pos.	noun noun_prop noun noun	noun noun_prop prep noun noun	0.95

TABLE 3 – Comparaison entre une formule et une séquence sur chaque couche d'information linguistique. Nous indiquons pour chaque couche le score de similarité cosinus obtenu.

4.3 Résultats

Séquence	Translittération	Translation	Score	Fréq.
غضب غضباً شديد	ġḍb ġḍban šdīd	il se mit dans une grande colère	0.89	7
وغضب غضبا شديد	w-ġḍb ġḍbā šdīd	et il se mit dans une grande colère	0.89	2
عرنوس غضبا شديد	<f-ġḍb> ʿrnūs ġḍbā šdīd	ʿArnūs <se mit dans> une grande colère	0.81	1
وغضب غضباً شديد	w-ġḍb ġḍban šdīd	et il se mit dans une grande colère	0.78	6
فغضب غضباً شديد	f-ġḍb ġḍban šdīd	alors il se mit dans une grande colère	0.78	3
غضبان غضبا شديد	ġḍbān ġḍbā šdīd	il est dans une grande colère	0.74	1
وغضبت غضباً شديد	w-ġḍbt ġḍban šdīd	elle se mit dans une grande colère	0.63	1
احبك حباً شديد	aḥbk ḥban šdīd	je t'aime d'un amour fort	0.46	1
الاسلام قتالاً شديد	<w-qātlt> l-islām qtālan šdīd	<et> les musulmans <menèrent> un combat acharné	0.46	1
وفرح فرحاً شديد	w-frḥ frḥan šdīd	et il se mit dans une grande joie	0.31	1

TABLE 4 – Exemples de Séquences similaires à « غضب غضباً شديد » (il se mit dans une grande colère).

Le Tableau 4 présente un échantillon du classement obtenu pour la formule « غضب غضباً شديد » (il se mit dans une grande colère). Nous avons isolé et classé 20 386 séquences, dont 7 329 avec une similarité cosinus supérieure à 0,5. La Figure 1 représente la distribution de nos séquences en fonction de leurs scores. Nous remarquons que seulement 813 séquences dépassent un seuil de 0,7.

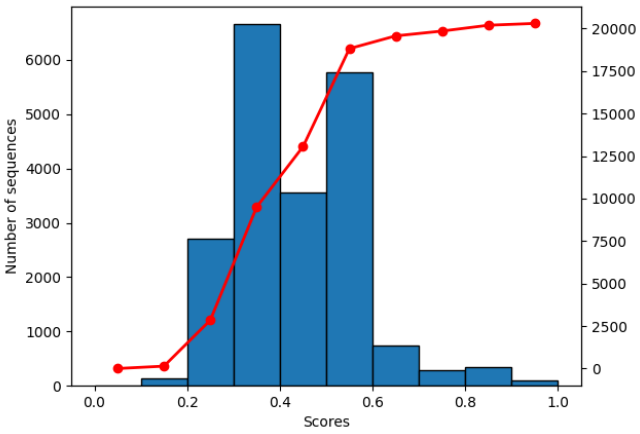

FIGURE 1 – Distribution de toutes les séquences isolées selon leur score. La ligne rouge correspond au nombre cumulé de séquences extraites de notre corpus.

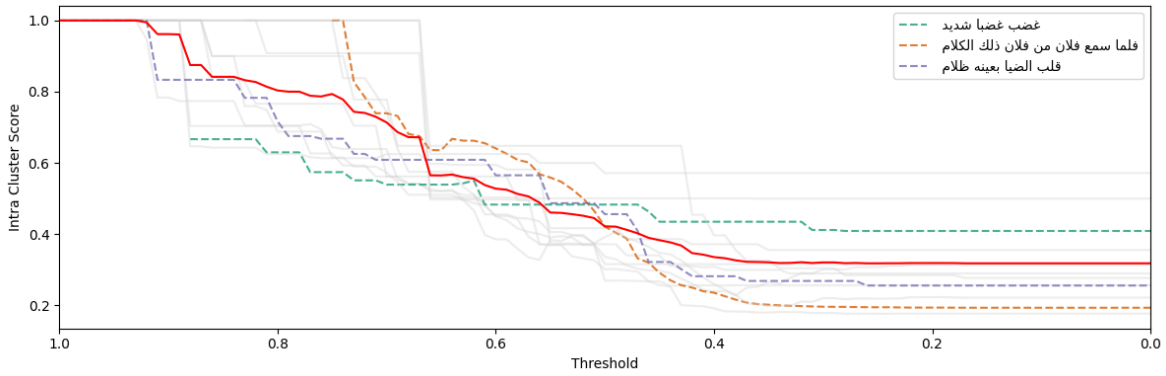

FIGURE 2 – Progression du score intra-cluster (ordonnée) pour chaque formule selon X (abscisse). La ligne rouge correspond à la moyenne de ce score pour toutes les formules.

Afin d'évaluer la qualité de notre classement, nous calculons un score intra-cluster, correspondant à la moyenne des similarités de la liste de séquences. Pour chaque formule, nous calculons le score intra-cluster de toutes les séquences avec un score $\geq X$, avec X variant de 0 à 1. La progression du score intra-cluster en fonction du seuil est présenté dans la Figure 2. Nous observons sans surprise que le score intra-cluster diminue quand X diminue. Nous proposons un étude approfondie des clusters en Annexes (Figures 3 et 4).

5 Discussion

Les variations relevées apparaissent principalement à trois niveaux. Au niveau graphique, certaines lettres et diacritiques ne sont pas toujours présentes. Par exemple, la *hamza* dans *ḍyāʾ* (3) est absente la plupart du temps, bien qu'elle soit indiquée dans d'autres variantes. Ce trait avait déjà été repéré par Jérôme Lentin (Lentin, 2008) : « la *hamza* finale est généralement absente »[6]. Cette fluidité graphique est aussi visible dans la préposition *fī*, tantôt écrite avec les deux points du *yāʾ* tantôt sans,

6. "final hamza is generally absent" (Lentin, 2008)

mais aussi sur la voyelle du *tanwīn* dans la formule (1), qui n'est pas systématiquement indiquée. Sur le plan morphologique, un des éléments les plus variables est le verbe, qui peut être conjugué à toutes les personnes, genres et nombres, comme dans la formule (2) où *smʿ* dépend du sujet et peut aussi bien se trouver sous la forme *smʿt* que celle de *smʿū*. C'est également le cas dans la formule (3) où l'on trouve *ġḍb* aussi bien que *ġḍbū* et *ġḍbt*. Nous avons trouvé plusieurs variantes de la formule (3) avec le duel *ʿynīh* à la place du singulier *ʿynh* (voir Section 6). Enfin, les variations se produisent au niveau lexical et affectent le verbe ou le nom, remplacés par un synonyme ou une image très proche de celle de la formule d'origine. Dans la formule (3), *ʿyn* (oeil) devient *wjh* (visage) ; on trouve aussi *ṣār* (devenir) à la place de *qlb* (se transformer), de même que *ġḍb* (se mettre en colère) à la place de *frḥ* (se mettre en joie) dans la formule (1). Pour ces deux derniers exemples, on pourrait avancer que les deux lexèmes ne sont pas synonymes. En réalité, ils appartiennent toujours au même champ lexical, celui des émotions : ils ne modifient pas le sens profond de la formule, et il nous semble que l'effet de repère produit par celle-ci (expliqué en Section 3.2) reste intact.

Toutefois, certaines variations modifient complètement le sens de la formule, au point d'en constituer une nouvelle. Ainsi, si *smʿ* (entendre) et *fhm* (comprendre), présent dans l'une des variantes de la formule (2), sont substituables, c'est parce que leur signification est très proche. En revanche, la variante avec *frġ mn* dans « فلما فرغ من ذلك الكلام » (et lorsqu'il eut exprimé ces paroles) semble être une formule à part entière. Nous avons remarqué que la formule (2) est systématiquement utilisée dans un contexte de dialogue, après qu'un personnage a tenu des propos en affectant un autre. Or, la variante avec le verbe *frġ mn* a son propre contexte d'utilisation : elle est employée après une déclamation de vers. Dans cette optique, *ġḍb* peut être remplacé par *frḥ* dans la formule (1), car les deux dénotent une émotion forte ressentie par le personnage. En revanche, la variante avec *qātl* « وقاتلت الاسلام قتالاً شديد » (et les musulmans menèrent un combat acharné) ne fait pas référence à une émotion, et donne une autre signification à la formule. Elle a son propre contexte d'utilisation dans la SĪRA, qui est celui des scènes de bataille. Ces exemples montrent que la variation linguistique dans ce corpus et dans les formules ne se produit pas de manière aléatoire. Certaines variantes dépassent le seuil de compréhension de la formule, ce qui indique qu'elles ne sont plus des variantes mais plutôt qu'elles constituent une autre formule. Le fait qu'elles aient leur propre contexte d'utilisation vient appuyer cet argument.

Certains éléments sont strictement invariables sur les plans syntaxique et sémantique. La structure syntaxique de la formule reste intacte : dans la formule (1), malgré toutes les variations possibles, la présence du *mafʿul muṭlaq* est constante et n'admet aucun type de variation. Nous pouvons aussi remarquer qu'il y a, dans chacune des formules, au moins un mot à position fixe qui ne semble changer sur aucun plan - graphique, morphologique ou lexical - et qui est rarement utilisé en dehors du style formulaire. Par exemple, on trouve 75 occurrences de l'adjectif *šdīd* dans la formule (1), et seulement 8 en dehors de ce contexte. De la même manière, on trouve 47 occurrences du nom *ẓlām* dans la formule (3), et seulement 5 en dehors. Les formules suivent aussi un schéma sémantique : la formule (2) a son propre contexte d'apparition qui ne peut être remplacé sans modifier un élément important de la formule (comme *smʿ* qui devient *frġ*). Toutes les variations de la formule (1) décrivent un sentiment fort : la colère (*ġḍb*), la joie (*frḥ*), l'amour (*ḥbb*) ou le tourment (*ʿḏb*). Lorsque la variation lexicale dépasse le sens, comme c'est le cas pour la variante avec *qātl* (tuer), le sens de la formule n'est pas atteint, et ce tournant sémantique provoque un processus de défigement, comme défini par (Mejri, 2009). Bien que nous n'ayons trouvé aucune variante qui souligne un processus de défigement dans la formule (3), on peut penser qu'une variation lexicale qui implique un tournant sémantique ne sera pas considérée comme faisant partie de la même formule.

6 Conclusion

Nous avons présenté une méthode d'identification et d'extraction de formules et de leurs variantes dans un corpus en moyen arabe. Nous avons extrait 20 386 séquences plus ou moins proches de ces formules. Ces séquences ont ensuite été classées en calculant une similarité entre plusieurs couches d'informations linguistiques. Au total, 813 séquences avec un score supérieur ou égal à 0,7 ont été trouvées. Nous avons également proposé une analyse qualitative des variantes obtenues pour trois formules. Cette analyse révèle que les variations se produisent principalement aux niveaux lexical, morphologique et graphique, mais jamais aux niveaux syntaxique et sémantique. De fait, toute variation affectant ces deux derniers niveaux forme en réalité une formule d'un nouveau type. Pour améliorer la chaîne de traitement présentée, nous envisageons de développer des outils de TAL dédiés au moyen arabe. Nous prévoyons également l'annotation d'une partie des séquences extraites pour analyser plus finement les performances de notre approche.

Remerciements

Nous tenons à remercier le programme ANR LiPoL [ANR 19-CE27-0024] qui a rendu ce corpus disponible dans le cadre de notre étude.

Références

ABDELALI A., DARWISH K., DURRANI N. & MUBARAK H. (2016). Farasa : A fast and furious segmenter for Arabic. In J. DENERO, M. FINLAYSON & S. REDDY, Éds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, p. 11–16, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-3003.

AL-SHARGI F., KAPLAN A., ESKANDER R., HABASH N. & RAMBOW O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1300–1306, Portorož, Slovenia : European Language Resources Association (ELRA).

ALHARBI R., MAGDY W., DARWISH K., ABDELALI A. & MUBARAK H. (2018). Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

ANONYMOUS (2000–2022). *Sī*rat al-Malik al-Ẓāhir Baybarṣ ḥasab al-riwāya al-šāmiyya*. éd. G. Bohas, S. Diab, I. Hassan, K. Zakharia, Damas and Beyrouth, Presses de l'Ifpo.

BEZANÇON J. & LEJEUNE G. (2023). Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels. In C. SERVAN & A. VILNAT, Éds., *30e*

Conférence sur le Traitement Automatique des Langues Naturelles, TALN, p. 56–67, Paris, France : ATALA.

BLAU J. (1982). The state of research in the field of the linguistic study of middle arabic. In *Études de Linguistique Arabe*, p. 187–203. Brill.

CONSTANT M., ERYIĞIT G., MONTI J., VAN DER PLAS L., RAMISCH C., ROSNER M. & TODIRASCU A. (2017). Survey : Multiword Expression Processing : A Survey. *Computational Linguistics*, **43**(4), 837–892. Place : Cambridge, MA Publisher : MIT Press, DOI : 10.1162/COLI_a_00302.

DARWISH K., MUBARAK H., ABDELALI A., ELDESOUKI M., SAMIH Y., ALHARBI R., ATTIA M., MAGDY W. & KALLMEYER L. (2018). Multi-Dialect Arabic POS Tagging : A CRF Approach. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

FERGUSON C. A. (1959). Diglossia. *WORD*, **15**(2), 325–340. DOI : 10.1080/00437956.1959.11659702.

GUILLAUME J.-P. (2004). Les scènes de bataille dans le roman de baybars : considérations sur le" style formulaire" dans la tradition épique arabe. *Arabica*, p. 55–76.

HABASH N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers.

KATZ G. & DIAB M. (2011). Introduction to the special issue on arabic computational linguistics. *ACM Transactions on Asian Language Information Processing*, **10**(1). DOI : 10.1145/1929908.1929909.

KHALIFA S., HASSAN S. & HABASH N. (2017). A morphological analyzer for Gulf Arabic verbs. In N. HABASH, M. DIAB, K. DARWISH, W. EL-HAJJ, H. AL-KHALIFA, H. BOUAMOR, N. TOMEH, M. EL-HAJ & W. ZAGHOUANI, Éds., *Proceedings of the Third Arabic Natural Language Processing Workshop*, p. 35–45, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1305.

LARCHER P. (2001). Moyen arabe et arabe moyen. *Arabica*, **48**(Fasc. 4), 578–609.

LENTIN J. (2003). Variétés d'arabe dans des manuscrits syriens du roman de baybars et histoire du texte. *Jean-Claude Garcin (sous la direction de) & Gisèle Seimandi (eds.), Lectures du Roman de Baybars, Marseille : Parenthèses-MMSH*, p. 91–111.

LENTIN J. (2008). *Middle Arabic*. Volume 3 de (Versteegh *et al.*, 2008).

LENTIN J. (2012). Reflections on middle arabic. *High vs Low and Mixed Varieties : Domain, Status and Function across Time and Languages, edited by Gunvor Mejdell and Edzard Lutz*, p. 32–51.

MEJRI S. (2009). Figement, défigement et traduction. Problématique théorique. In *figement, défigement et traduction (Fijación, desautomatización y traducción)*, p. 153–163. Universidad de Alicante.

OBEID O., INOUE G. & HABASH N. (2022). Camelira : An Arabic multi-dialect morphological disambiguator. In W. CHE & E. SHUTOVA, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 319–326, Abu Dhabi, UAE : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-demos.32.

PARRY M. (1930). Studies in the epic technique of oral verse-making. i. homer and homeric style. *Harvard Studies in Classical Philology*, **41**, 73–147.

PASHA A., AL-BADRASHINY M., DIAB M. T., EL KHOLY A., ESKANDER R., HABASH N., POOLEERY M., RAMBOW O. & ROTH R. (2014). Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec*, volume 14, p. 1094–1101.

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2002). Multiword Expressions : A Pain in the Neck for NLP. In A. GELBUKH, Éd., *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, p. 1–15, Berlin, Heidelberg : Springer. DOI : 10.1007/3-540-45715-1_1.

SAMIH Y., ATTIA M., ELDESOUKI M., ABDELALI A., MUBARAK H., KALLMEYER L. & DARWISH K. (2017). A neural architecture for dialectal Arabic segmentation. In N. HABASH, M. DIAB, K. DARWISH, W. EL-HAJJ, H. AL-KHALIFA, H. BOUAMOR, N. TOMEH, M. EL-HAJ & W. ZAGHOUANI, Éds., *Proceedings of the Third Arabic Natural Language Processing Workshop*, p. 46–54, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1306.

VERSTEEGH K. *et al.* (2008). *Encyclopedia of Arabic Language and Linguistics*, volume 3. Brill.

ZACK L. & SCHIPPERS A. (2012). *Middle Arabic and Mixed Arabic : Diachrony and Synchrony*. Leiden, The Netherlands : Brill. DOI : 10.1163/9789004228047.

ZALMOUT N., ERDMANN A. & HABASH N. (2018). Noise-robust morphological disambiguation for dialectal Arabic. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 953–964, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1087.

Annexes

Nous présentons ici trois Tableaux additionnels : le Tableau 5 présente les 13 formules sur lesquelles nous avons travaillé. Le Tableau 6 ainsi que le Tableau 7 sont deux classements additionnels, respectivement pour les formules « قلب الضيا بعينه ظلام » et « فلما سمع فلان من فلان ذلك الكلام ». Finalement, les Figures 4 et 3 comprennent les distributions de nos segments extraits dans un espace vectoriel par rapport à la formule ayant servi à leur extraction.

À propos de la fonctionnalité Levantine de CAMELIRA : nous n'avons pas eu l'occasion d'utiliser cette fonctionnalité, qui n'était pas disponible au moment de nos expériences. Cependant, cette même fonctionnalité a été rendue disponible quelques jours avant la date limite de la soumission. Nous n'avons donc pas eu le temps de l'implémenter dans ce travail.

Concernant la translittération : sauf exceptions, le corpus de SĪRAT BAYBARṢ n'est pas vocalisé. Nous ne disposons d'aucun enregistrement ni d'aucun témoignage de lecture à haute voix. De ce fait, nous utilisons le système de translittération employé par d'autres chercheurs sur le moyen arabe, et qui consiste à ne pas présumer des voyelles brèves, car nous n'avons aucune indication sur la manière dont elles étaient prononcées dans une variété d'arabe aussi mixte. Par exemple, « غضب » translitteré *ġaḍiba* pour des textes standards, est dans cet article translitteré *ġḍb*.

En ce qui concerne le procédé de vectorisation de nos phrases, séquences et formules, nous utilisons la fonction *CountVectorizer* de la librairie python SCI-KIT LEARN avec les paramètres suivants :

$ngram\_range = (1, 1)$	$encoding = "utf - 8"$	$lowercase = True$
$stop\_words = None$	$analyzer = lambda\ x : x.split(" ")$	

Formule	Translittération	Translation
قلب الضيا بعينه ظلام	qlb aḍ-ḍyā bʿynh ẓlām	la lumière dans son oeil se transforma en ténèbres
غضب غضباً شديد	ġḍb ġḍbā šdīd	il se mit très en colère
فلما سمع فلان من فلان ذلك الكلام	f-lmmā smʿ flān mn flān ḏlk al-klām	lorsque A eut entendu de la part de B ces paroles
فز واثب على الاقدام	fz wāṯb ʿlā al-aqdām	il sauta en bondissant sur ses pieds
بات ذلك الليله	bāt ḏlk al-lylh	il dormit cette nuit-là
اصبح الصباح	aṣbḥ aṣ-ṣbāḥ	le jour se leva
اظلم الظلام	aẓlm aẓ-ẓlām	la nuit s'assombrit
دقت طبول الانفصال	dqqt ṭbūl al-anfṣāl	les tambours de la séparation résonnèrent
وعند فراغه من ذلك الكلام	w-ʿnd frāġh mn ḏlk al-klām	quand il eut exprimé ces paroles
اما سمعت ما قال الشاعر	amā smʿt mā qāl aš-šāʿr	n'as-tu pas entendu ce que dit le poète
اما سمعت الشاعر حيث قال	amā smʿt aš-šāʿr ḥyṯ qāl	n'as-tu pas entendu le poète lorsqu'il dit
وأنشد وقال	w-ʾnšd w-qāl	il dit en déclamant
أشاد يسجع نفسه بهذه الأبيات	ʾšād ysjʿ nfsh b-hḏh al-ʾbyat	il éleva la voix, faisant rimer ces vers

TABLE 5 – Les 13 formules sur lesquelles nous fondons nos travaux.

Séquence	Translittération	Translation	Score	Fréq.
قلب الضيا بعينه ظلام	qlb aḍ-ḍyā bʿynh ẓlām	la lumière dans son (m.) oeil se transforma en ténèbres	1.0	21
قلب الضيا بعينيه ظلام	qlb aḍ-ḍyā bʿynīh ẓlām	la lumière dans ses (m.) yeux se transforma en ténèbres	0.92	6
قلب الضيا بعينها ظلام	qlb aḍ-ḍyā bʿynhā ẓlām	la lumière dans ses (f.) yeux se transforma en ténèbres	0.92	2
قلب الضيا فى عينيه ظلام	qlb aḍ-ḍyā fī ʿynīh ẓlām	la lumière en ses (m.) yeux se transforma en ténèbres	0.84	1
صار الضيا بعينه ظلام	ṣār aḍ-ḍyā bʿynh ẓlām	la lumière dans ses (m.) yeux devint ténèbres	0.8	2
قلب الضيا بعينه طلام	qlb aḍ-ḍyā bʿynh ṭlām	la lumière dans son (m.) oeil se transforma en ténèbres	0.8	1
قلب الضيا فى وجهه ظلام	qlb aḍ-ḍyā fī wjhh ẓlām	la lumière en son visage se transforma en ténèbres	0.77	1
قلب الضياء بعينيه ظلام	qlb aḍ-ḍyāʾbʿynīh ẓlām	la lumière dans ses (m.) yeux se transforma en ténèbres	0.73	1
صار الضيا بعينيه ظلام	ṣār aḍ-ḍyā bʿynīh ẓlām	la lumière dans ses (m.) yeux devint ténèbres	0.72	1

TABLE 6 – Quelques séquences isolées avec la formule « قلب الضيا بعينه ظلام » (the light in his eyes turned into darkness).

Séquence	Translittération	Translation	Score	Fréq.
فلما سمع الملك من القاضى ذلك الكلام	f-lmmā smʿ al-mlk mn al-qāḍī ḏlk al-klām	lorsque le roi eut entendu de la part du qāḍī ces paroles	0.75	1
فلما سمع الملك منه ذلك الكلام	f-lmmā smʿ al-mlk mnh ḏlk al-klām	lorsque le roi eut entendu de sa part ces paroles	0.74	3
فلما سمع الملك من ابراهيم ذلك الكلام	f-lmmā smʿ al-mlk mn brāhīm ḏlk al-klām	lorsque le roi eut entendu de la part d'Ibrahim ces paroles	0.73	2
فلما سمع الملك من عماد ذلك الكلام	f-lmmā smʿ al-mlk mn ʿmād ḏlk al-klām	lorsque le roi eut entendu de la part de ʿImād ces paroles	0.73	1
فلما سمع الملك من عيسى ذلك الكلام	f-lmmā smʿ al-mlk mn ʿysā ḏlk al-klām	lorsque le roi eut entendu de la part de ʿIssa ces paroles	0.73	1
فلما سمع ذلك الكلام	f-lmmā smʿ ḏlk al-klām	lorsqu'il eut entendu ces paroles	0.72	6
فلما سمع الملك ذلك الكلام	f-lmmā smʿ al-mlk ḏlk al-klām	lorsque le roi eut entendu ces paroles	0.71	44
فلما سمع عرنوس ذلك الكلام	f-lmmā smʿ ʿrnūs ḏlk al-klām	lorsque ʿArnous eut entendu ces paroles	0.71	11
فلما فرغ من ذلك الكلام	f-lmmā frġ mn ḏlk al-klām	lorsqu'il eut exprimé ces paroles	0.69	1
فلما فهم الملك ذلك الكلام	f-lmmā fhm al-mlk ḏlk al-klām	lorsque le roi eut compris ces paroles	0.58	4

TABLE 7 – Quelques séquences isolées avec la formule « فلما سمع فلان من فلان ذلك الكلام » (when A heard those words from B).

Pour chaque formule, nous projetons les vecteurs de toutes ses séquences dans un espace vectoriel à deux dimensions (Figures 3 et 4). Plus le score d'une séquence est élevé, plus son point correspondant tend vers le rouge. Au contraire, plus son score est bas, plus ce même point tend vers le bleu. La formule est représentée par un point noir. Les séquences possédant des scores plus élevés semblent se

regrouper autour de la formule dans l'espace vectoriel. Nous proposons dans la section suivante une analyse qualitative effectuée sur un nombre limité de formules.

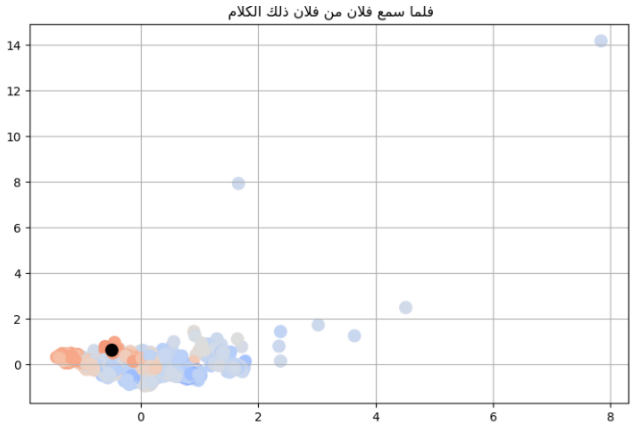

FIGURE 3 – Distribution des séquences isolées pour la formule « فلما سمع فلان من فلان ذلك الكلام » dans un espace vectoriel à deux dimensions. Cette formule présente des séquences atypiques, très éloignées de la formule recherchée. De ce fait, nous la séparons des autres formules.

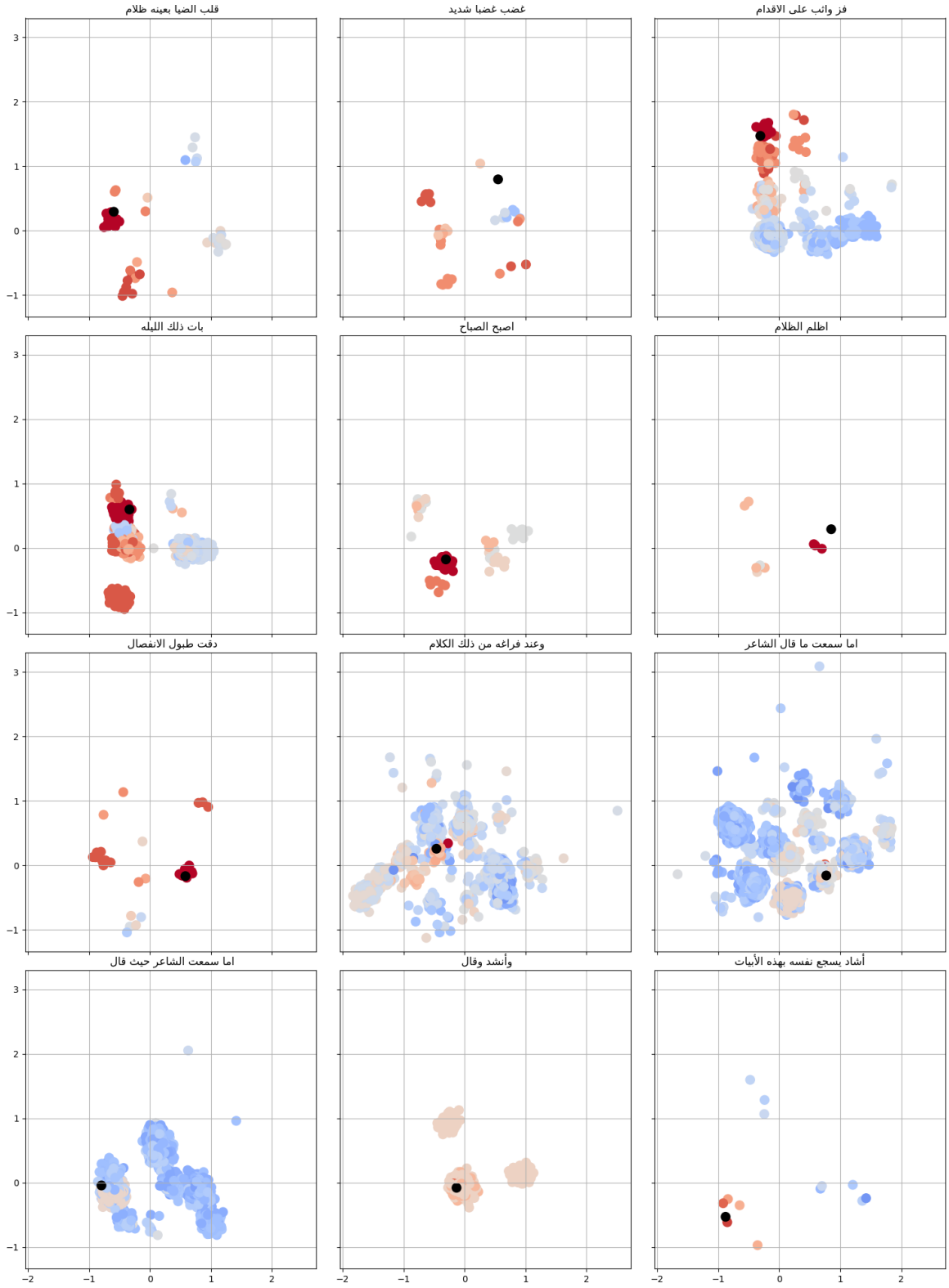

FIGURE 4 – Distribution des séquences isolées pour chaque formule dans des espaces vectoriels à deux dimensions. Les points rouges indiquent des séquences possédant un score élevé, tandis que les points bleus correspondent aux séquences avec un score bas. Les points noirs représentent nos formules.