Lost in Variation : An Unsupervised Methodology for Mining Lexico-syntactic Patterns in Middle Arabic Texts

Julien Bezançon
STIH, CERES,
Sorbonne Université

Rimane Karam
Orient & Méditerranée, CERES,
Sorbonne Université
LiPoL, Ifpo

Gaël Lejeune
STIH, CERES,
Sorbonne Université

28 rue Serpente 75006 Paris, France
firstname.lastname@sorbonne-universite.fr

Abstract

Although Modern Standard Arabic and some dialects of Arabic have been extensively studied in NLP, Middle Arabic is still very much unknown to the field. However, Middle Arabic presents challenges not addressed by current NLP tools. In particular, it is characterized by variation since it mixes standard features, colloquial ones, as well as features that belong to neither of the two. Here, we introduce a methodology to identify, extract and classify variations of 13 manually retrieved 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 and composed of 53,843 sentences. In total, 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.

1 Introduction

As described in Guellil et al. (2019), three main types of Arabic have been covered by NLP research: Classical Arabic, Modern Standard Arabic (MSA) and dialects (Egyptian, Gulf, ...). While Classical Arabic has been the subject of only a few works, MSA and dialects have been the focus of a fair number of studies. This is not the case for Middle Arabic, which, to the best of our knowledge, has not been studied from a NLP perspective.

However, Middle Arabic study in NLP is interesting on its own. Middle Arabic is "distinguished by its linguistically (and therefore stylistically) mixed nature, as it combines standard and colloquial features with others of a third type, neither standard nor colloquial" (Lentin, 2008). As a result, Middle Arabic texts tend to have a wide range of possible variations for a given structure (Zack and Schippers, 2012). By studying Middle Arabic in NLP, we would be able to produce and process new resources which take into account numerous varieties of Arabic simultaneously. This would be useful for better understanding text processing in Arabic, as Arabic texts are rarely written with a single variety of Arabic (Katz and Diab, 2011).

Studying a corpus of Middle Arabic can be challenging for both linguists and NLP experts, being of mixed nature and prone to variation, as discussed in Section 2.1. For instance, formulas like "فز واثب على الاقدام" ("he leaped jumping on his feet") can also be written as "نهض واثب على الاقدام" ("he got up jumping on his feet") or with the graphical variation "فذ واثب على الاقدام" (*fḏḏ* instead of *fzz*). This challenge is compounded by other difficulties specific to Arabic processing in NLP, including orthographic ambiguity, morphological richness and orthographic noise (Habash, 2010).

Here, we aim to provide a new methodology to study a corpus with multiple varieties of Arabic. Our goal is the identification of all possible variations for a given formula. To do so, we introduce a corpus of Middle Arabic, SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ, composed of 53,843 sentences along with 13 formulas that were manually retrieved by a linguist expert and whose variations we want to study. We plan to use token alignment techniques, lexico-syntactic patterns as well as similarity measures in order to extract and rank each possible variation of a given formula.

We find that our study is similar to the ones dealing with multiword expressions (MWEs) in NLP. MWEs are generally seen as conventionalized and idiomatic sequences (Sag et al., 2002). In MWE processing, the identification task, whose goal is to identify MWEs in a text, shares a lot of similarities with the work we try to achieve (Constant et al., 2017). For this reason, we plan to use the methodology presented in (Bezançon and Lejeune, 2023),

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 25–37

created for the identification and the extraction of MWEs and unfrozen MWEs, i.e. MWEs which have undergone lexical, syntactic and/or semantic changes.

We first introduce the notion of Middle Arabic in Section 2. We then introduce the corpus and the formulas we used to test our methodology in Section 3. Those formulas correspond to short and frequently occurring instances in our corpus. Hereafter, we show the different steps that led to the identification and extraction of those formulas and their look-alikes in Section 4. Finally, we discuss the variations we observe between the original formulas and their newly-found variations in Section 5.

2 Middle Arabic: a non standardized language

2.1 Definition

Arabic is usually perceived as a two-sided language: standard on the one hand, and colloquial on the other. This linguistic situation, called diglossia, was widely theorized by Charles Ferguson: the "high" variety refers to the standard, whereas the "low" stands for the dialects (Ferguson, 1959). The linguistic reality of Arabic is actually not as binary and hermetical, and Ferguson himself acknowledged the existence of intermediate varieties. Further research has defined these varieties that lay between the two poles of diglossia under the term Middle Arabic (Blau, 1982). Middle Arabic can thus be described as a set of intermediate registers that mixes both standard and colloquial features, and also has features of its own, that are not standard nor colloquial and that belong to a third pole (Larcher, 2001).

A whole area of Arabic literature has been written in Middle Arabic, and it was shown that it had nothing to do with poor language skills in *fuṣḥā* (Classical Arabic). We have examples of texts written by the same scholars both in perfect *fuṣḥā* and in Middle Arabic; and popular literature is, for a large part, written in some varieties of Middle Arabic, just as the THOUSAND AND ONE NIGHTS (Lentin, 2012). The Damascene version of SĪRAT BAYBARṢ, which we work on, is another example, and one should keep in mind that even though the text seems close to Levantine dialects, not only does it have standard features, but it also has very specific features that belong to neither of the two poles. For instance, the relative pronoun *alladṯ* in its masculine singular form remains invariable regardless of the gender and number of its antecedent (Lentin, 2012).

Thus, Middle Arabic can be distinguished by its mixed nature: it combines features from both standard and dialects. Given this situation, it makes it complex to use either standard or dialect tools such as part-of-speech taggers on a Middle Arabic corpus. Middle Arabic being a mixed, hybrid set of varieties of Arabic that tends to play on the linguistic continuum, it creates an important amount of linguistic variation throughout the text. Isolating manually all the variations of the same formulas in our corpus can be difficult given the language of the text and the size of the corpus. A closer look into Arabic NLP research could help us develop an automatic approach on Middle Arabic texts that might be expanded to other languages with frequent variations.

2.2 NLP Tools and Resources

MSA and Dialects Arabic studies are potentially the most useful for this work as explained in Section 2.1. On the one hand, there is a wide variety of tools used to process data in MSA, like segmenters (Abdelali et al., 2016) and morphosyntactic taggers (Zalmout et al., 2018; Pasha et al., 2014). On the other, we can find tools specific to each dialect, like Egyptian (Zalmout et al., 2018; Samih et al., 2017) or Gulf (Alharbi et al., 2018; Khalifa et al., 2017), but there are also tools that can handle several dialects simultaneously (Darwish et al., 2018; Al-Shargi et al., 2016).

To our knowledge, there are no NLP tools dedicated to Middle Arabic. This can be a problem for a language marked by linguistic variation such as Middle Arabic, that has standard as well as dialect features, but also features of its own that are neither of the two. Faced with this challenge, we plan to use CAMELTOOLS (Obeid et al., 2022) as a substitute to label our corpus in Section 4.1. It is a multi-dialect morphological disambiguation tool covering MSA as well as Egyptian, Gulf, and Levantine. While it is unlikely that this tool will identify and tag correctly Middle Arabic features, we suppose that tagging MSA and dialectal Arabic ones is at reach.

Vol.	# Tokens	# Sent.	T/S	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: Statistics for each volume (**Vol.**) of the SĪRAT BAYBARṢ corpus. In addition to the number of tokens and sentences (**Sent.**), we give the average sentence size in tokens (**T/S**) and the Type Token Ration (**TTR**).

3 Dataset description

3.1 Corpus

SĪRAT AL-MALIK AL-ẒĀHIR BAYBARṢ is a popular prose epic cycle from the Ottoman period. It is a text designed above all for performance since it used to be told by *hakawātī*-s, storytellers of the Levant, who memorized and recited the stories in cafés or homes, by heart or with the help of booklets. For this project, we are using the Damascene version of SĪRAT BAYBARṢ (Anonymous, 2000–2021). This composite corpus consists of a set of booklets of manuscripts written down by many different scribes between the 18th and the 20th century, and gathered afterwards by three storytellers from Damascus [1]. We decided to focus on the first 90 booklets of the Abu Ahmad manuscript, named after the storyteller who compiled it. It is normally composed of 183 booklets, but only the first 90 have been digitalized so far. In the edited version (Anonymous, 2000–2021), they have been segmented into 9 volumes of 10 booklets each. Table 1 shows various statistics for each volume. We notice that the Type-to-Token Ratio (TTR) is very low for the whole corpus (7.06 %), which can indicate that a lot of constructions are repeated over and over.

Another particularity should be noted regarding the language of SĪRAT BAYBARṢ, in addition to it being mostly close to the Damascene dialect. Some characters are made fun of and portrayed as caricatures in their way of speaking, either because they come from another country or because they represent the enemy. These two layers of variation combined - Middle Arabic and idiolects within the SĪRA - complexify any kind of statistics on this text, especially given the absence of tools to explore Middle Arabic to our knowledge.

3.2 Formulas

We are looking for sequences within the SĪRAT BAYBARṢ that occur regularly in specific contexts. As shown by the work of J. P. Guillaume which is very close to ours linguistic-wise (Guillaume, 2004), each occurrence of a given sequence bears the same meaning despite the linguistic variations, without denoting a narrative progression in the story. For instance, these sequences can indicate a sudden change of a character's mood, or be used as opening or closing sequences in a situation, whether it is a new day dawning, the night falling, a poem declamation or even battle scenes for example. The formulation, regularity and context of these sequences make them easy to be noticed by the reader (or the listener, in a performance situation) regardless of the variations. As described in Section 2.1, Middle Arabic is characterized by linguistic variation, and these sequences are no exception. In a way, they are similar to MWEs, as they are conventionalized in our corpus and tend to have similar, almost fixed forms.

The works of Milman Parry on the Homeric style could help define these sequences. Parry described "a group of words which is regularly employed under the same metrical conditions to express a given essential idea" under the term formula (Parry, 1930). His corpus of reference is the Homeric poems, a versified text. Although it has come down to us written, it is deeply rooted in the oral tradition. As we said before in Section 3.1, orality is an important element in our corpus as it was also destined to a performance. As for the versification part, we can argue that although written in prose, our corpus is still punctuated by sequences that have a very close usage to Homer's "as soon as early rosy-fingered Dawn appeared" for "when it was morning". Moreover, these formulas in SĪRAT BAYBARṢ happen to be used in the context of *saj*ʿ (rhymed prose). They do not follow versification rules, but they do not strictly belong to prose either, especially because they tend to provoke other rhyming formulas in a row. Despite the lack of versification constraints, we can assume that other types of constraints, either linguistic or stylistic, impact the formulas in SĪRAT BAYBARṢ. Their core concept consists in their regularity and in the importance of expressing an idea

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

"without second thought" (Parry, 1930), which fits our corpus. Formulas are a landmark, for the poet / scribe as well as for the listener / reader, and their presence in the text with so many variations might tell something about the language. We aim to see how these variations occur within a formula, with the hypothesis that they do not happen randomly but that they rather follow some pattern.

Thirteen frequently occurring sequences were found by a linguist expert who is also very familiar with the SĪRAT BAYBARṢ corpus. Those sequences correspond to formulas in our corpus. We base our experiment on them. For intelligibility reasons, we chose to present three of them in order to give detailed results and examples:

1. غضب غضباً شديد
   (*ġḍb ġḍban šdīd*)
   "he got very angry"
2. لما سمع فلان من فلان ذلك الكلام
   (*lmmā smʿ flān mn flān ḏlk al-klām*)
   "when A heard those words from B"
3. قلب الضيا بعينه ظلام
   (*qlb aḍ-ḍyā b-ʿynh ẓlām*)
   "the light in his eye turned into darkness"

(1) denotes a very strong feeling (namely anger) resulting from a situation or an action taken by another character. (2) appears very often after a character has said something that affected another character, whatever type of impact it is (positive or negative), which leads most of the time to an action by the latter character or a sudden change of mood. (3) denotes a sudden and abrupt change of mood, resulting often from what a character has just said. In fact, the last two formulas frequently follow one another. Our goal is, for each formula, to automatically find similar sequences that exhibit only slight variations. For instance, for (1), we aim to find similar sequences like (a.), (b.) and (c.).

a. غضبوا غضباً شديد
   (*ġḍb**ū** ġḍban šdīd*)
   "they got very angry"

b. غضبان غضبا شديد
   (***ġḍbān** ġḍb**ā** šdīd*)
   "he is very angry"

c. وفرح فرحاً شديد
   (***w-frḥ frḥan** šdīd*)
   "and he got very happy"

sentence:	". شديد حباً احبك انا ، والله : لى فقال"
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: Example of an entry of the SĪRAT BAYBARṢ corpus.

These three variants give an idea of what types of variation are possible within the same formula. They can be morphological and impact the verb such as *ġḍbū* in (a.) in place of *ġḍb* in the original formula. The variations can also be graphic and guide the presence or absence of some letters or diacritics, such as in (b.). The double vowel marker of the *tanwīn* (nunation, i.e. the mark of indefiniteness) is absent in *ġḍbā* even though the *ʾalif* is written, whereas (a.) indicates it in *ġḍban*. Finally, these variations can occur at a lexical level, changing completely the lexeme while preserving the structure of the sequence, as shown in (c.) where the verb *ġḍb* used in a. and b. (to get angry) becomes *frḥ* (to get happy).

4 Methodology

4.1 Processing Middle Arabic

We used CAMELTOOLS (Obeid et al., 2022) to (i) tokenize the corpus, (ii) get POS tags, (iii) get lemmas and (iv) segment it into sentences. Table 2 shows an entry of the corpus. The scripts we used to process the SĪRAT BAYBARṢ corpus are available in a dedicated GitHub repository[2]. CAMELTOOLS was chosen for its ability to handle different dialects of Arabic. Indeed, most Arabic morphosyntactic taggers have been designed to annotate Modern Standard Arabic only, as stated by Obeid et al. (2022); Darwish et al. (2018). We could have tried to use CAMELTOOLS's Levantine tool in conjunction with standard Arabic tools, to cover both the Damascene and the standard features of our corpus. Unfortunately, except for the online demo version of CAMELTOOLS, which only allows us to enter very few words in the input bar, the Levantine model was not available.

[2] https://github.com/JulienBez/ASMR

Layer	Formula	Sequence	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: Searched formula and found sequence side by side for each linguistic layer, with a cosine similarity score.

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

Table 4: Example of alignment at token level.

To roughly evaluate the quality of the annotation of CAMELTOOLS, we manually annotated 71 sentences for a total of 1,037 annotated tokens. Both the annotator and CAMELTOOLS had to choose between 5 tags for each token: noun, preposition, numeral, punctuation or verb. The precision of CAMELTOOLS on those tokens was 91.99 %, which can be considered low, since we drastically reduced the complexity of the tag set. The performance of the part-of-speech tagging and lemmatization is probably not as reliable as it would be for an MSA text. For instance, as table 2 shows, "شديد" (*šdīd*) is analysed as a "noun_prop" whereas it is an adjective. However, we did not expect perfect results, and we think it provides a basis that will be useful for different purposes.

4.2 Sequence Association

Our first step was to associate each sentence of our corpus with the formulas it resembles. We did a fuzzy matching between each sentence of our corpus and each of the manually chosen formulas by creating vectors and calculating cosine distance scores. If the distance between a formula and a sentence was too high (> to 0.9), we didn't associate them. By doing so, we only associate sentences and formulas with a minimum commonality of elements. For instance, the sentence "غضب غضبا شديد" ("he got very angry") was associated with the formula "انا احبك حباً شديد" ("I love you very much") with a cosine distance score of 0.87. Additionally, each sentence can be associated with more than one formula.

4.3 Candidates Ranking

For each sentence, we want to know if it contains at least one of the formulas it has been associated with, *in extenso* or with slight variations. We adapted the code and methodology presented in (Bezançon and Lejeune, 2023) for Arabic. The author's goal was to find unfrozen multiword expressions, i.e. multiword expressions which have been modified to some degree (for instance "may the force be with you" becoming "may the peace be with you"). This methodology was created to find both exact matches with a given sequence and closely related matches, i.e. matches that show a slight degree of variation and can therefore be linked to the original sequence. In the remainder of this subsection, we describe the different steps used to find and rank candidates based on their similarity to the formula they were associated with:

Alignment For each sentence, we aligned it with its associated formula to highlight their common tokens. As an example, we give in table 4 the alignment between the sentence "قلب الضيا فى وجهه ظلام" ("the light in his face turned into darkness") and the formula "قلب الضيا بعينه ظلام" ("the light in his eye turned into darkness"). This alignment shows us that the word "بعينه" has been replaced by the words "فى وجهه" in the sentence. Those alignments were made with BIOPYTHON[3], as this package's alignment process proposes all possible alignments between two sequences.

Segmentation We used the alignments to isolate common sequences between a sentence and a formula. Those sequences correspond to the longest subsequence of words that begin and end with the same words with a minimal number of misalignments (i.e. the minimal edit distance at token level). For instance, in the alignment presented in table 4, the complete sentence would be isolated since it (i) begins with the same word as the formula (قلب) and (ii) end with the same word (ظلام). A sentence can have more than one sequence with a formula.

[3] https://biopython.org/

Sequence	Transliteration	Translation	Score	Freq
غضب غضباً شديد	ġḍb ġḍban šdīd	he got very angry	0.89	7
وغضب غضبا شديد	w-ġḍb ġḍbā šdīd	and he got very angry	0.89	2
عرنوس غضبا شديد	<f-ġḍb> ʿrnūs ġḍbā šdīd	so ʿArnūs <got> very angry	0.81	1
وغضب غضباً شديد	w-ġḍb ġḍban šdīd	and he got very angry	0.78	6
فغضب غضباً شديد	f-ġḍb ġḍban šdīd	so he got very angry	0.78	3
غضبان غضبا شديد	ġḍbān ġḍbā šdīd	he is very angry	0.74	1
وغضبت غضباً شديد	w-ġḍbt ġḍban šdīd	and she got very angry	0.63	1
احبك حباً شديد	aḥbk ḥban šdīd	I love you very much	0.46	1
الاسلام قتالاً شديد	<w-qātlt> l-islām qtālan šdīd	<and> the muslims <fought> very hard	0.46	1
وفرح فرحاً شديد	w-frḥ frḥan šdīd	and he got very happy	0.31	1

Table 5: Some ranked sequence candidates for the formula "غضب غضباً شديد" ("he got very angry"). We show sequences with a high score as well as sequences with a lower similarity score on purpose.

Similarity Measurement We vectorized each sequence with its associated formula before calculating a cosine similarity score between them. The higher the score, the closer the sequence and the formula tend to be. Therefore, a score of 1 indicates a perfect match, while a score of 0 informs us that there is no common element between them. This measure is performed at different levels, as shown in the next paragraph.

Ranking We ranked each sequence according to its similarity with the formula. This ranking relies on several linguistic features (tokens, POS tags and lemmas) by calculating an average score from the cosine similarity obtained with each feature. Thus, the alignment, segmentation and measurement steps were repeated for every additional linguistic feature. Table 3 shows the sequence "قلب الضيا فى وجهه ظلام" ("the light in his face turned into darkness") compared with its associated formula with respect to the different linguistic layers we spoke off.

4.4 Results

The results take the form of a ranking for each formula we searched for. Table 5 shows the ranking obtained for the formula "غضب غضباً شديد" (1). In total, we found and ranked 20,386 sequences, including 7,329 with a cosine similarity above 0.5. Figure 1 shows the distribution of found sequences according to their score. We find that the higher the score, the fewer the corresponding sequences. Thus, only 813 sequences have a score of 0.7 or more. To evaluate the quality of our ranking, we used an intra cluster similarity score. This score

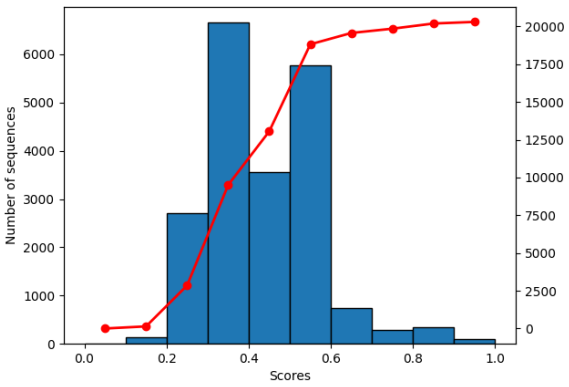

Figure 1: Distribution of found sequences according to their score for every formula we searched for. The red line shows the cumulative number of sequences found.

is obtained by computing the mean of a cosine similarity matrix created from a list of sequences s, as shown in Equation 1. The higher the intra-cluster score, the closer the sequences.

$$s_1, s_2, ..., s_n \Rightarrow \begin{pmatrix} s_1.s_1 & s_1.s_2 & ... & s_1.s_n \\ ... & ... & ... & ... \\ s_n.s_1 & s_n.s_2 & ... & s_n.s_n \end{pmatrix} \quad (1)$$

For each formula, we calculated the intra-cluster score of every sequence related to it with a score $\geq X$, X being equal to 1. We progressively lowered X to include more sequences from our ranking and to calculate the progression of the intra-cluster scores. Figure 2 is the result of this process. We observe that the lower X, the lower the intra cluster score. This fact could indicate that, for a given formula, our ranking seems to put the most

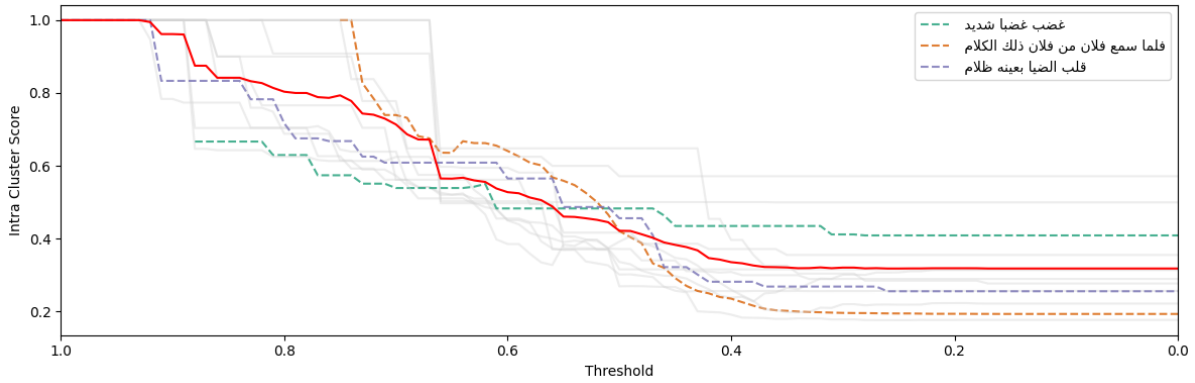

Figure 2: Progression of the intra cluster score (y-axis) for each formula, according to X (x-axis). The doted lines represent the formulas we focused on in this paper. Other formulas are shown in gray. The red line is the mean intra-cluster score for every formula.

similar sequences to this formula at the top while putting the less relevant ones at the bottom. We also mapped the vectors of every sequence found for each formula in a two-dimensional vectorial space. Figure 3 shows three formulas as an example. Sequences with a high score are represented by red dots, while sequences with a low score are represented by blue dots. The formulas are represented by a black dot. We observe that sequences with a high score tend to be closer to formulas than sequences with a low score. In the remainder of this paper, we propose an analysis of the results we obtained for a selected set of formulas.

5 Discussion

The variations we observed appear mainly on three levels : graphical, morphological and lexical. On the graphical level, we noticed that some letters and diacritics are not always indicated. For instance, the *hamza* in *ḍyāʾ* (3) is most of the time absent, despite it being written in some variants of the same formula. This feature was already described by Lentin in (Lentin, 2008) : "final *hamza* is generally absent". This graphical flexibility is also visible within the preposition *fī*, sometimes written without the two points of the *yā*ʾ, as well as the double vowels of the *tanwīn* in (1) which are not systematically indicated. On a morphological level, one of the most variable elements is the verb, which can be conjugated at any person and in any number or gender, as in (2) where *smʿ* depends on the subject, and can become *smʿt* or *smʿū* . It is also the case in (3) where *ġḍb* can be *ġḍbū* as well as *ġḍbt*. We also found many variants of 3 with *ʿynīh* in the dual form instead of *ʿynh* (see 7). Finally, variations can occur on a lexical level, either on verbs or nouns that are synonyms or describe a very close image, preserving the meaning of the formula. In (3) *ʿyn* ("eye") becomes *wjh* ("face"), and *qlb* ("turned to") can be replaced by *ṣār* ("became"), as well as *ġḍb* ("to get angry") by *frḥ* ("to get happy") in (1). For the last two, one could argue that they are not synonyms. In fact, as we will show in the next paragraphs, they still belong to the same lexical field (emotions, for instance): they do not affect the core meaning of the formula, and the landmark effect that we explained in 3.2 is preserved.

Nevertheless, some of these variations do change completely the meaning of the formula, to the point of consisting of another formula. For instance, if *smʿ* (to hear) and *fhm* (to understand), a verb that we found in one of the variants of the formula (2), are exchangeable, it is because the meaning of the two verbs - at least in this context - is very close; whereas the variant with *frġ mn* in "فلما فرغ من ذلك الكلام" ("and when he had said those words") affects way too much the formula, and thus consists in another formula. Indeed, we noticed that the formula (2) is systematically used in the context of dialog, right after one character has said something that affects another character. In opposition, the same formula with *frġ* has its own specific context: it is only used after a character has recited a poem. Following the same logic, *ġḍb* can be substituted by *frḥ* in (1), because both indicate emotions or feelings that overwhelm a character. The variant with *qātl* "وقاتلت الاسلام قتالاً شديد" ("the muslims fought very hard"), which does not denote an emotion, gives another meaning to the

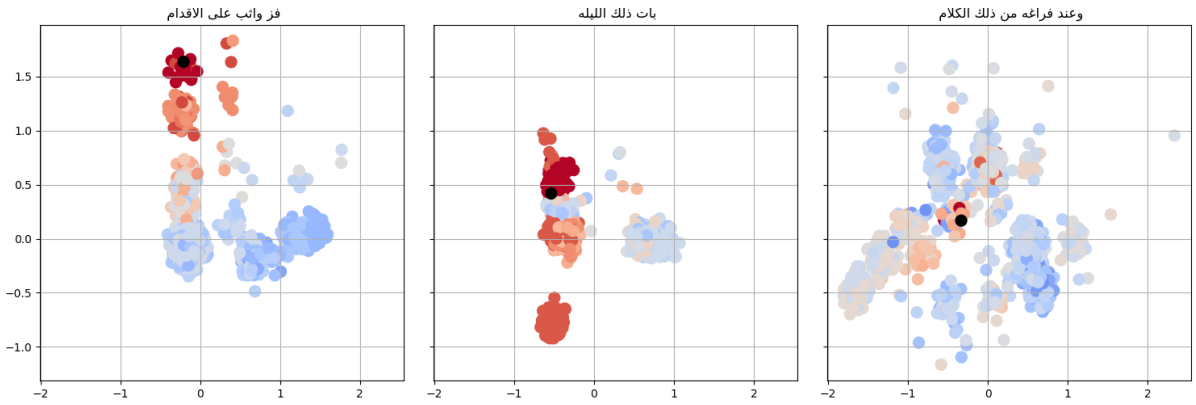

Figure 3: Distribution of found sequences for three formulas on a two-dimensional vectorial space. Red dots correspond to sequences with high scores in our ranking, while blue dots correspond to sequences with low scores. Black dots represent our formulas.

formula. In fact, it has its own context of use, which is the battle scenes that happen within the SĪRA. All these examples show that linguistic variation in our corpus and within the specific context of the formulas do not occur randomly. Some variants pass the threshold of comprehension, which indicates that they no longer belong to the same formula. The fact that they have their own context of use supports this idea.

In fact, some elements are strictly invariable, and they all happen to be syntactic and semantic. The syntactic structure of the formula stays unchanged: in (1), the *mafᶜul muṭlaq* structure is constant in all the variants of the formula, regardless of any graphical, lexical or morphological changes. We can also note that there is at least one static word in each formula: a word that never changes graphically, morphologically or lexically, with a fixed position in the formula, and which is hardly ever used in unrelated found sequences. For instance, *šdīd* occurs 75 times within the formula (1), and only 8 outside of it; *ẓlām* has 47 occurrences within the formula (3), and only 5 outside of it. The formulas also follow a semantic pattern: as we explained in the previous paragraph, (2) has a specific context of occurrence which cannot be replaced by another without changing a strong parameter in the formula (as when *smᶜ* becomes *frġ*). In (1), regardless of the lexical changes, all the variants of the formula describe a very strong feeling, whether it is anger (*ġḍb*), joy (*frḥ*), love (*ḥbb*) or torment (*ᶜḏb*). When the lexical variation exceeds this meaning, as in the variant with *qātl* ("to kill"), the semantic level is not reached, and this meaning shift leads to an unfreezing process, as defined by (Mejri, 2009). Although we did not find any variant that underlines an unfreezing process in formula (3), such as in (1) and (2), we can guess that any lexical variation that involves a meaning shift will not be considered as part of the same formula.

6 Conclusion

In this paper, we presented a methodology for the identification and extraction of formulas likely to be subject to variations in a Middle Arabic corpus. We extracted 20,386 sequences resembling these formulas. We ranked those sequences according to their similarity with the searched formulas on various linguistic layers. In total, 813 segments with a score of 0.7 or more were found.

This process helped us get an overview of the variants of each formula. We noticed that some elements of a formula can easily vary whereas others are strictly invariable. Variations may occur at the lexical, morphological and graphical level but never on a syntactic nor semantic level. If any kind of variation happens on the last two levels mentioned, it changes completely the essence of the formula, consisting in another formula of another type which is used in its own specific context.

In future work, we aim to build improved NLP tools for processing Middle Arabic. It would help us to analyze more formulas, than the set we studied in this paper. We also plan to work with linguists experts in Damascene in order to annotate a sample of the sequences found. This would help us to propose further analysis of the performances of the methodology we presented here. We hope this work will be helpful for further research on non-standard Arabic variants.

Acknowledgements

We would like to thank the ANR LiPoL program [ANR 19-CE27-0024] for making this corpus available.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California. Association for Computational Linguistics.

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1300–1306, Portorož, Slovenia. European Language Resources Association (ELRA).

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed AbdelAli, and Hamdy Mubarak. 2018. Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anonymous. 2000–2021. *Sīrat 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.

Julien Bezançon and Gaël Lejeune. 2023. Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels. In *30e Conférence sur le Traitement Automatique des Langues Naturelles, TALN*, pages 56–67, Paris, France. ATALA.

Joshua Blau. 1982. The state of research in the field of the linguistic study of middle arabic. In *Études de Linguistique Arabe*, pages 187–203. Brill.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892. Place: Cambridge, MA Publisher: MIT Press.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-Dialect Arabic POS Tagging: A CRF Approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2019. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Jean-Patrick Guillaume. 2004. Les scènes de bataille dans le roman de baybars: considérations sur le" style formulaire" dans la tradition épique arabe. *Arabica*, pages 55–76.

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

Graham Katz and Mona Diab. 2011. Introduction to the special issue on arabic computational linguistics. *ACM Transactions on Asian Language Information Processing*, 10(1).

Salam Khalifa, Sara Hassan, and Nizar Habash. 2017. A morphological analyzer for Gulf Arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 35–45, Valencia, Spain. Association for Computational Linguistics.

Pierre Larcher. 2001. Moyen arabe et arabe moyen. *Arabica*, 48(Fasc. 4):578–609.

Jérôme Lentin. 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*, pages 32–51.

Jérôme Lentin. 2008. *Middle Arabic*. Volume 3 of (Versteegh et al., 2008).

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

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

Milman Parry. 1930. Studies in the epic technique of oral verse-making. i. homer and homeric style. *Harvard Studies in Classical Philology*, 41:73–147.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec*, volume 14, pages 1094–1101.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg. Springer.

Younes Samih, Mohammed Attia, Mohamed Eldesouki, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer, and Kareem Darwish. 2017. A neural architecture for dialectal Arabic segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 46–54, Valencia, Spain. Association for Computational Linguistics.

Kees Versteegh et al. 2008. *Encyclopedia of Arabic Language and Linguistics*, volume 3. Brill.

Liesbeth Zack and Arie Schippers. 2012. *Middle Arabic and Mixed Arabic: Diachrony and Synchrony*. Brill, Leiden, The Netherlands.

Nasser Zalmout, Alexander Erdmann, and Nizar Habash. 2018. Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 953–964, New Orleans, Louisiana. Association for Computational Linguistics.

Appendix

In this Appendix, we show 3 additional Tables. Table 6 shows the 13 formulas we based our work on. Tables 7 and 8 are two additional ranking for the formulas "قلب الضيا بعينه ظلام" and "فلما سمع فلان من فلان ذلك الكلام". Figure 4 and 5 shows the distribution of found sequences for all formulas on a two-dimensional vectorial space.

About the Levantine feature of CAMELIRA: The Levantine module was made available a few days before the conference deadline. We therefore did not have the opportunity to use it in this work.

About the transliteration: the SĪRAT BAYBARṢ corpus is not vocalized (with a few rare exceptions) and we have no record nor any kind of testimony on how the text was read aloud. Therefore, we chose to follow the transliteration system used by other researchers on Middle Arabic, which consists of not assuming the short vowels, because we simply do not know and have no indication on how they were supposed to be pronounced in such a mixed variety of Arabic. For instance, the world "غضب", transliterated as *ġaḍiba* for standard texts, is transliterated as *ġḍb* in the present paper.

For our experiment, we used *sci-kit learn*'s vectorization features with the following parameters:

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

Formula	Transliteration	Translation
قلب الضيا بعينه ظلام	qlb aḍ-ḍyā bʿynh ẓlām	the light in his eye turned into darkness
غضب غضباً شديد	ġḍb ġḍbā šdīd	he got very angry
فلما سمع فلان من فلان ذلك الكلام	f-lmmā smʿ flān mn flān ḏlk al-klām	when A heard from B those words
فز واثب على الاقدام	fz wāṯb ʿlā al-aqdām	he leaped jumping on his feet
بات ذلك الليله	bāt ḏlk al-lylh	he slept that night
اصبح الصباح	aṣbḥ aṣ-ṣbāḥ	it became morning
اظلم الظلام	aẓlm aẓ-ẓlām	it became night
دقت طبول الانفصال	dqqt ṭbūl al-anfṣāl	the drums of separation rumbled
وعند فراغه من ذلك الكلام	w-ʿnd frāġh mn ḏlk al-klām	when he had said those words
اما سمعت ما قال الشاعر	amā smʿt mā qāl aš-šāʿr	haven't you heard what the poet said
اما سمعت الشاعر حيث قال	amā smʿt aš-šāʿr ḥyṯ qāl	haven't you heard the poet when he said
وأنشد وقال	w-ʾnšd w-qāl	he chanted and said
أشاد يسجع نفسه بهذه الأبيات	ʾšād ysjʿ nfsh b-hḏh al-ʾbyat	he praised, rhyming himself with these verses

Table 6: The 13 formulas we based our work on.

Sequence	Transliteration	Translation	Score	Freq
قلب الضيا بعينه ظلام	qlb aḍ-ḍyā bʿynh ẓlām	the light in his eye turned into darkness	1.0	21
قلب الضيا بعينيه ظلام	qlb aḍ-ḍyā bʿynīh ẓlām	the light in his eyes turned into darkness	0.92	6
قلب الضيا بعينيه ظلام	qlb aḍ-ḍyā bʿynīh ẓlām	the light in his eyes turned into darkness	0.92	6
قلب الضيا بعينها ظلام	qlb aḍ-ḍyā bʿynhā ẓlām	the light in her eye turned into darkness	0.92	2
قلب الضيا فى عينيه ظلام	qlb aḍ-ḍyā fī ʿynīh ẓlām	the light in his eyes turned into darkness	0.84	1
صار الضيا بعينه ظلام	ṣār aḍ-ḍyā bʿynh ẓlām	the light in his eye became darkness	0.8	2
قلب الضيا بعينه طلام	qlb aḍ-ḍyā bʿynh ṭlām	the light in his eye turned into darkness	0.8	1
قلب الضيا فى وجهه ظلام	qlb aḍ-ḍyā fī wjhh ẓlām	the light in his face turned into darkness	0.77	1
قلب الضياء بعينيه ظلام	qlb aḍ-ḍyāʾbʿynīh ẓlām	the light in his eyes turned into darkness	0.73	1
صار الضيا بعينيه ظلام	ṣār aḍ-ḍyā bʿynīh ẓlām	the light in his eyes became darkness	0.72	1

Table 7: Some ranked sequences for the formula "قلب الضيا بعينه ظلام" ("the light in his eyes turned into darkness").

Sequence	Transliteration	Translation	Score	Freq
فلما سمع الملك من القاضى ذلك الكلام	f-lmmā smʿ al-mlk mn al-qāḍī ḏlk al-klām	when the king heard from the qāḍī those words	0.75	1
فلما سمع الملك منه ذلك الكلام	f-lmmā smʿ al-mlk mnh ḏlk al-klām	when the king heard from him those words	0.74	3
فلما سمع الملك من ابراهيم ذلك الكلام	f-lmmā smʿ al-mlk mn brāhīm ḏlk al-klām	when the king heard from Ibrahim those words	0.73	2
فلما سمع الملك من عماد ذلك الكلام	f-lmmā smʿ al-mlk mn ʿmād ḏlk al-klām	when the king heard from ʿImad those words	0.73	1
فلما سمع الملك من عيسى ذلك الكلام	f-lmmā smʿ al-mlk mn ʿysā ḏlk al-klām	when the king heard from ʿIssa those words	0.73	1
فلما سمع ذلك الكلام	f-lmmā smʿ ḏlk al-klām	when he heard those words	0.72	6
فلما سمع الملك ذلك الكلام	f-lmmā smʿ al-mlk ḏlk al-klām	when the king heard those words	0.71	44
فلما سمع عرنوس ذلك الكلام	f-lmmā smʿ ʿrnūs ḏlk al-klām	when the king heard from ʿrnus those words	0.71	11
فلما فرغ من ذلك الكلام	f-lmmā frg mn ḏlk al-klām	when he had said those words	0.69	1
فلما فهم الملك ذلك الكلام	f-lmmā fhm al-mlk ḏlk al-klām	when the king understood those words	0.58	4

Table 8: Some ranked sequences for the formula "فلما سمع فلان من فلان ذلك الكلام" ("when A heard those words from B").

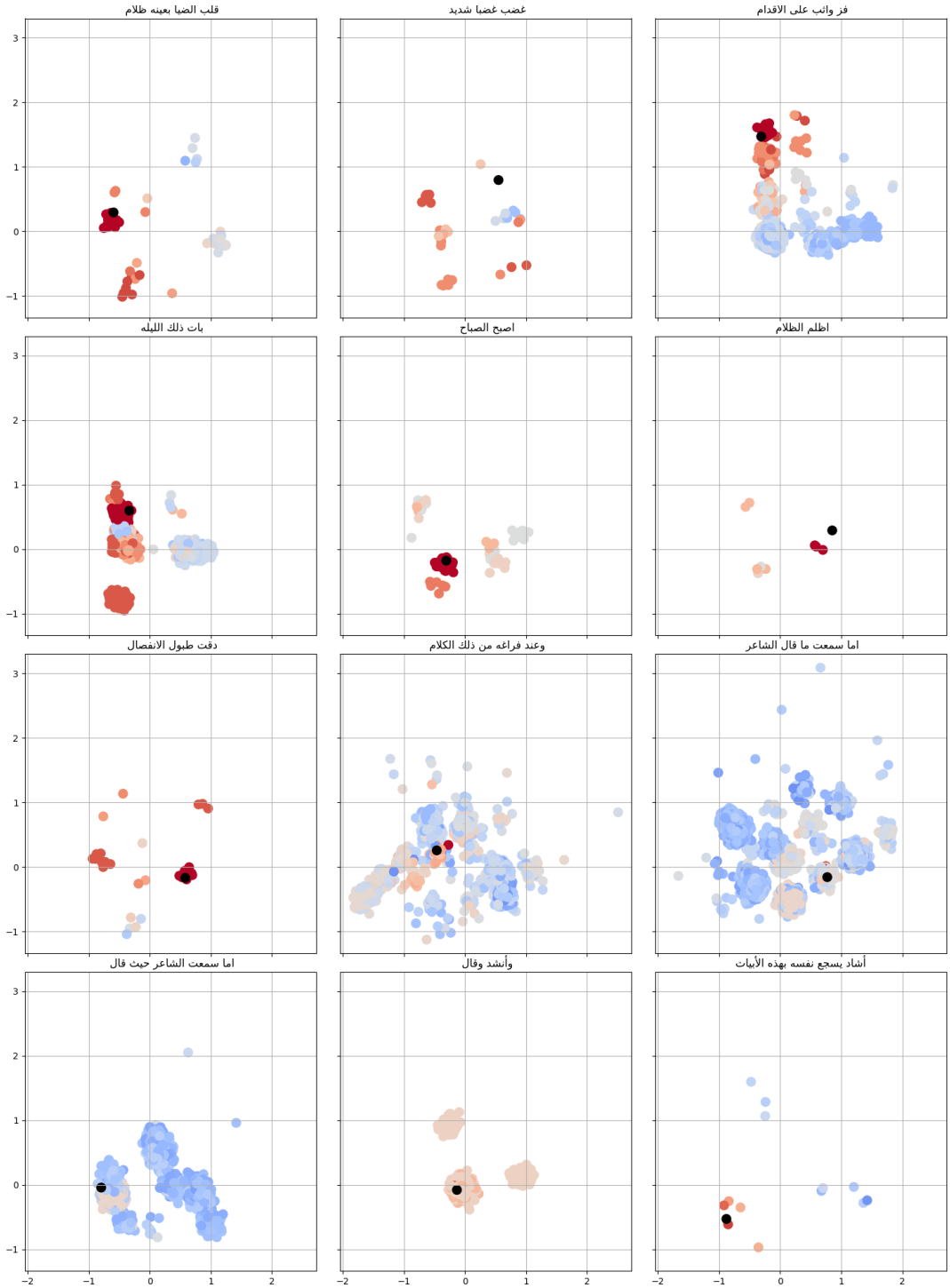

Figure 4: Distribution of found sequences for all formulas on a two-dimensional vectorial space. Red dots correspond to sequences with high scores in our ranking, while blue dots correspond to sequences with low scores. Black dots represent our formulas.

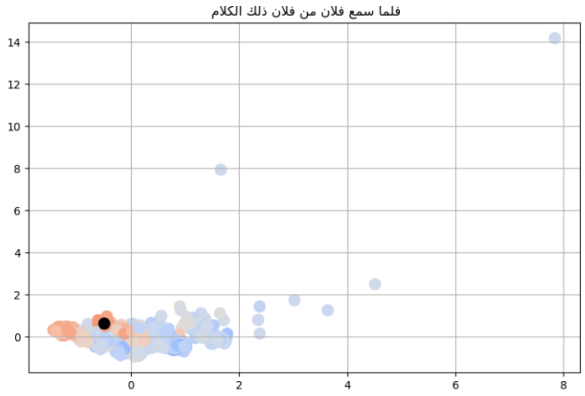

Figure 5: Distribution of found sequences for the formula فلما سمع فلان من فلان ذلك الكلام on a two-dimensional vectorial space. We separated this formula from the others because it has significant outliers.