Decoding Byzantine Book Epigrams: an Exploration of Machine-assisted Extraction of Formulaic Material

Kyriaki Giannikou†, Colin Swaelens‡, Ilse De Vos⋆, Els Lefever‡, Klaas Bentein†
†Dpt. of Linguistics, ‡Language & Translation Technology Team, ⋆Flanders AI Academy
†Blandijnberg 2, Ghent, Belgium ‡Groot-Brittannielaan 45, Ghent, Belgium
⋆Kasteelpark Arenberg 10, Leuven, Belgium
{author1, author2, author4, author5}@ugent.be
author3@kuleuven.be

Abstract

This paper proposes a machine-assisted methodology for identifying and extracting formulaic sequences from a subset of the Database of Byzantine Book Epigrams (DBBE). The methodology involves conceptualising formulaicity within the DBBE corpus, pre-processing and extracting n-grams from textual data, followed by refinement before delving into the interpretation of the results. Through systematic application of this methodology, some initial insights into the characteristics of formulaic language within the Byzantine book epigram tradition are gained. Representative findings illustrate the nature of recurring patterns, cases of creative elaboration, and their content. This initial exploration aims to facilitate a deeper understanding of the concept of formulaicity in Byzantine book epigrams; while computational analysis provides a quantitative perspective, linguistic and philological research is necessary for a more nuanced understanding. Future research directions include refining the methodology and expanding the scope of analysis beyond the current subset of the DBBE. Overall, this study lays the groundwork for further research on this rich book epigram tradition.

Keywords: Byzantine Greek, Corpus Linguistics, Natural Language Processing

1. Introduction

Formulaic language plays a pivotal role in linguistic communication across various domains, from TV shows to religious rituals, or regal ceremonies; formulaicity is omnipresent. It is characterised by recurrent sequences of varying degrees of fixedness and compositionality, ranging from collocations to idioms. Its study traverses various linguistic disciplines, while it has also captured the interest of scholars in literary studies.

This paper delves into the rich tapestry of formulaic language within the context of the Database of Byzantine Book Epigrams (DBBE) corpus, a repository teeming with paratextual material that offers invaluable sociolinguistic insights into Byzantine society's book culture (Ricceri et al., 2023). A Byzantine book epigram is defined as a poem, a text in verse, found in the margins of Byzantine manuscripts; it is written in a book and refers to the book's production and consumption (Kominis, 1966; Lauxtermann et al., 2003; Bernard and Demoen, 2019). In general, book epigrams are written by scribes who are copying the main text of a manuscript and provide the reader with more insights into that particular manuscript. In Example 1, found at the end of a manuscript, the scribe breathes a sigh of relief that the copying of the text is finished. DBBE Occurrence 17571 is just one attestation, however, of a whole series comprising over 200 epigrams. These vary in length, ranging from one to four verses, and are characterised by the employment of different words and different word orders.

Given that some book epigrams are attested over a hundred times, being copied, adapted, or extended by other scribes, it is argued here that book epigrams establish a fully-fledged literary genre. This copying results in repetition of half, one or multiple verses, or even complete epigrams, which brings us to the hypothesis that this corpus of book epigrams displays noticeable formulaicity. Although this claim is an admitted truth (Bernard and Demoen, 2019; Ricceri et al., 2023), a comprehensive linguistic study of the formulaic and creative aspects of book epigrams has only recently begun. This paper marks the initial step of this endeavour: the machine-assisted extraction and identification of formulaic material from this particular corpus.

(1) Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα,
οὕτως καὶ οἱ γράφοντες βιβλίου τέλος.
Hōsper xenoi chairousin idein patrida
houtōs kai hoi graphontes bibliou telos.
Just like travellers rejoice upon seeing their homeland, so do writers at the end of a book.[1]
DBBE Occurrence 17571

Our exploration is anchored in a multidisciplinary approach, drawing from computational methodologies, historical linguistics, and literary conventions to unearth the formulaic material embedded within these epigrams.

[1]All translations are provided by the authors.

Our undertaking begins with an examination of relevant literature (Section 2). A detailed description of the DBBE data and our selected subset follows in Section 3. The method to extract and identify formulaic sequences is given in Section 4. Our methodology comprises a series of systematic steps, from data export and pre-processing to n-gram extraction and dictionary creation, leading to a manual evaluation of the results of the proposed machine-assisted method. This is followed by a discussion of the results (Section 5) and a conclusion with a glimpse in future research (Section 6) on capturing the formulaic language employed by scribes.

2. Literature Review

2.1. Formulaicity

Linguistic research on the phenomenon of formulaicity spans across various disciplines, including corpus linguistics, psycholinguistics, cognitive linguistics, and sociolinguistics. The initial systematic description of this phenomenon can be traced back to the 19th century and the observations of standardised speech production among aphasic patients. Subsequent contemporary studies have placed significant emphasis on the social dimension of formulaic language. While the exact terminology may vary among theorists, linguists generally agree on a continuum of formulaicity, with formulaic sequences ranging from collocations to idioms. This continuum is based on the degree of lexical, morphological and syntactic fixedness (i.e. the extent of variation permitted) and compositionality (i.e. whether the formula can be deconstructed into smaller, meaningful components).

Formulaicity is widely recognised as an inherent characteristic of language. Sinclair (1991) highlighted fixed linguistic sequences or *collocations* as prevalent in English corpora. As outlined in his *idiom* and *open-choice principles*, he attributed this prevalence to cognitive efficiency. There he also proposed a computational method for extracting collocations from large corpora, a methodology that has been successfully applied to various contemporary languages, e.g. the Bantu languages (de Schryver, 2008).

Wray (2002, 2008) has made significant contributions to the field. Her Morpheme Equivalent Units (MEUs) model suggested that prefabricated sequences are both stored and retrieved as a whole from the mental lexicon. This occurs unless there is a need for parsing, according to the Needs-Only Analysis (NOA) model she proposed. These prefabricated 'chunks' can be either *fully-fixed* or semi-preconstructed (*partially-fixed frames*). *Partially-fixed frames* can either allow more specific slots (i.e. very detailed variation) to be completed or less specific ones. An example of the former is variation in tense or mood of a verb, the latter is a placeholder that can hold any word at that specific slot. Formulaicity serves both social and cognitive functions in communication, according to Wray; it reduces processing effort for participants, aids in the manipulation of the hearer and contributes in marking discourse structure. Wray (2008) also discussed commonly used criteria for identifying formulaic sequences. These criteria include their frequency of occurrence, perceived formulaicity based on native speakers' intuition, phonological or spelling indicators of holistic processing or chunking, and fixedness – all of which, in her opinion, present weaknesses if used in isolation rather than in combination. Additionally, she proposed eleven diagnostic criteria for evaluating intuitive judgements.

Kuiper (2000) emphasised external, cultural factors, such as the need for socially accepted and predictable communication, along with the internal factor of memory constraints, as motivation for formulaic language use. Later, Kuiper (2009) defined formulae as 'lexical items with the features $[-\text{wlc} + \text{nlcu}]$'. This means that they cannot be broken down into meaningful sub-units, resulting in no Word-Level Complexity (wlc) – at least not retaining all their possible compositional meanings – and they do have specific Non-Linguistic Conditions of Use (nlcu), including 'speech community' and occasion. Most importantly, he introduced the concept of *formulaic genres* (e.g. weather forecasts), recognising that formulaic elements characterise larger discourse structures as formulaic in their own way, emphasising the interrelation of formulaic language with broader communicative genres and underscoring its role in discourse marking.

From a constructionist perspective, formulaic sequences are conceptualised as *form-function pairings*, i.e. constructions (Goldberg, 1995, 2006; Buerki, 2016b). The analysis of formulae as constructions, particularly of those that are more abstract, led Buerki (2016b) to regard the distinction between formulae and other constructions arbitrary. However, he still asserted that the psycholinguistic significance of formulaic sequences, indicated by Wray (2002), persists.

Research on formulaicity of course extends beyond English and includes a multitude of other modern languages; just to give some examples, formulaic sequences have been identified and extracted from Swedish (Cinková et al., 2006), Slovenian (Dobrovoljc, 2020), and Spanish (Cortes, 2022) corpora, using (semi-)automatic procedures. Buerki (2016a) provided a general guide to such procedures using software tools like N-Gram Processor[2]

[2] https://github.com/buerki/ngramprocessor

and Sub-String[3].

2.2. Historical Corpora

Recent linguistic research on formulaicity has broadened its scope to include historical corpora. For instance, Moulin et al. (2015) outlined the study of formulae in historical German (beginning of written tradition to the Early New High German period), while Rutten and van der Wal (2012) applied Wray's insights to a historical corpus of Dutch letters dating from the 17th and 18th centuries. The latter argued for the pertinence of formulaicity research, notwithstanding the non-contemporary and purely written nature of the corpus. Their findings indicated that formulaic language serves to compensate for limited literacy and writing experience within written contexts, just like it reduces processing effort within an oral context. This aligns with earlier observations that compared to more experienced speakers and performers, less experienced ones tend to rely more heavily on formulae, thereby enhancing their fluency (Lord, 1960; Leiwo, 2005; Bozzone, 2010). It should be noted that the particularities of their historical corpus necessitated adaptation; the recurrent references to the Christian god led to the establishment of a distinct *Christian-ritual formula*, complementing Wray's general communication-based functions of formulaic language. This showcases the feasibility of applying such a theoretical framework on a historical, written corpus, notwithstanding the peculiarities imposed by its socio-cultural context.

In the context of the Greek language, scholars have predominantly explored formulaicity and creative variation within literary corpora such as the Homeric poems. Parry and Lord's seminal work on the oral-formulaic theory (Lord, 1960) revealed the mnemonic function of recurring fixed expressions in traditional oral poetry. More recent scholarship (Bakker, 2005; Foley, 2007) has delved into the creative utilisation of formulae within the narratives of the Homeric epics. They have explored how these formulae are not static elements but rather dynamically employed, adapting to the oral tradition and the metrical constraints of the hexameter. Through their analyses, they have highlighted the fluidity and adaptability of formulae, shedding light on their pivotal role in shaping the rich tapestry of Homeric storytelling. Bozzone (2010, 2014), in particular, contributed to the discourse by analysing Homeric formulae as constructions, taking into consideration factors inhibiting formulaic reliance and poetic novelty. Jeffreys (1973), on the other hand, extended the Homeric formula theory to late Byzantine popular poetry, while later works (Jeffreys and Jeffreys, 1983, 1986) highlighted the impact of textual transmission on variation in formulae. Limited studies exist however on formulaic language in earlier Byzantine literary composition, and mostly discussing specific formulaic expressions, often not exhaustively (Garitte, 1962; Treu, 1977; Boeten et al., 2021), necessitating further investigation.

In recent years, there has been a notable shift in scholarly focus from prestigious literary texts to everyday documents, reflecting a broader interest in aspects of daily life. Non-literary (documentary) corpora, both ancient and medieval[4] Greek, are thus receiving more and more attention. The standardised, conventional language they present has led to systematic studies on their formulaic material, particularly within administrative texts, such as contracts, petitions, public inscriptions, and letters, originating from both private and business correspondence (Lazzarini, 1976; Nachtergaele, 2015; Bentein, 2023). The shift to everyday documents also sparked increased interest in paratextual material. Regarding the Byzantine period in particular, manuscripts have evolved beyond their initial perception as repositories of classical texts to being acknowledged as valuable reservoirs of paratextual material, a term coined by Genette (1987). Paratextual material includes elements accompanying the main text to provide information on, e.g., authorship, genre, and purpose. Scholars (Ciotti and Lin, 2016; Teeuwen and van Renswoude, 2018) now focus on paratexts to unveil historical sociolinguistic dimensions of manuscripts. They regard paratexts as vital markers that reveal the sociohistorical context of the manuscript and its intricate interplay with language, intellectuality, culture, and society.

The Database of Byzantine Book Epigrams project (Ricceri et al., 2023) has built, digitised, and made available such a paratextual corpus to the scholarly community, including scholars, linguists, and recently also Natural Language Processing (NLP) researchers and engineers. Byzantine book epigrams, written in the margins of manuscripts, intertwine poetic expression with practical details. As such, they shed light on facets such as the patrons of the manuscripts and the identities of the scribes involved in their transcription. They do so by reproducing or building upon standard, formulaic material, as it has been observed (Bernard and Demoen, 2019).

3. Data

The aim of the DBBE has been the digitisation and enrichment of a distinct textual corpus, specifically focusing on Byzantine book epigrams. Byzantine

[3] https://github.com/buerki/SubString

[4] Byzantine and medieval will be used within this paper as synonyms to address the period of the 5th until the 15th century.

book epigrams are paratextual in nature, as they exist on the threshold between the intellectual realm of the text and the physical manifestation of the book they are inscribed on. Overall, the corpus' significance is derived from its direct association with the material context of manuscripts, providing insights into codicology, palaeography, production, and reading practices in the context of Byzantine society. This makes them a valuable source for understanding the social dynamics of book culture. The DBBE corpus provides an excellent opportunity to shed light on the actors and communities involved in manuscript production and consumption and, therefore, on their formulaic and creative language use. Often originating from non-professional poets, these epigrams offer a glimpse into less erudite literary devices and linguistic developments.

3.1. The DBBE: Occurrences and Types

The organisation of book epigrams in the DBBE involves a categorisation of entries into 'Occurrences' and 'Types'. 'Occurrences' represent instances of epigrams as they appear in manuscripts, accommodating all variation in the original texts as transcribed by the DBBE team, based on the original manuscript and catalogues or related publications. 'Types', on the other hand, constitute reconstructed texts grouping similar 'Occurrences' while accommodating mostly minimal variation. Thus, 'Type' records present normalised, readable versions of the manuscript evidence. Although 'Types' provide scholars with a more homogeneous and standardised corpus in terms of language, this standardisation deprives us of access to the scribes' original linguistic choices and oversimplifies the dynamic nature of textual transmission. Book epigrams undergo copying and reworking, thus being transmitted in similar yet different forms, a phenomenon not clearly illustrated by the standardisation of the 'Types'. However, individual differences are preserved at the level of 'Occurrences' in any case. Table 1 shows that the DBBE holds 12,497 'Occurrences' that are all linked to one or more of the 5,022 'Types'.

	Epigrams	**Verses**	**Tokens**
Occurrences	12,497	48,458	272,426
Types	5,022	24,879	140,103
Scribe Types	1,849	13,150	32,888

Table 1: The number of epigrams, verses and tokens of both the DBBE Types and Occurrences.

Example 2 shows 'Type' 1974 2a and one of its corresponding 'Occurrences' in 2b. This 'Occurrence' is highly affected by itacism, the shift of the classical Athenian pronunciation of four vowels (ι *i*, η *è*, ε *e*, υ *u*) and two diphthongs (ει *ei*, οι *oi*) to one and the same [i] sound. The author of Example 2b clearly knew what he wanted to write, but did not stick to archaising orthographic conventions. That is why this 'Occurrence' is placed under the umbrella of Type 1974.

(2) a. Ἡ μὲν χειρ ἡ γράψασα σήπεται τάφῳ
Hè men kheir hè grapsasa sèpetai taphōi
Type 1974 v.1

b. Ἱ μὲν χὺρ ἡ γράψασα σύπτεται (!) τάφω
Hi men khyr hè grapsasa syptetai (!) taphō
Occurrence 18305 v.1

The hand that has written, is rotting in the grave

The rationale behind using DBBE 'Types' instead of 'Occurrences' for the presented research can be summarised as follows. First, at this preliminary stage, the primary objective is to develop a method for extracting recurring material to identify and define what constitutes a formula in the Byzantine Book Epigram corpus. Further exploration into mapping, analysing, and explaining the variation present in formulae will occur in subsequent research phases. Moreover, in the absence of a reliable lemmatisation tool, a degree of standardisation is essential for the initial automatic extraction of collocations, due to the significant morphological and orthographic variation present in Byzantine Greek corpora (cf. Example 2). Furthermore, when considering formulaic sequences at the level of 'Occurrences,' they often vary depending on their contextual information. This contextual information dictates the specific content of 'slots' that need to be filled between the standard, recurring, formulaic elements; e.g. when signing at the end of the work: χειρ (ADJ) NAME, 'the hand of (ADJ) NAME' (Example 3). The specific content of these slots is not relevant at this stage of formulaicity research, thus an examination at the level of 'Types' is more appropriate. Lastly, by extracting formulaic material from the 'Types,' we ensure that their frequency of occurrence reflects their usage in various contexts. Thus, the extracted sequences are more likely to constitute formulaic entities themselves, rather than formulations that, for example, were repeatedly copied as part of the exact same epigram. [5]

[5] It is important to note that, while this is mostly true, it is not guaranteed that if a formula appears in two distinct 'Types', the variation between these 'Types' will always be pertinent to the usage of that sequence. As this paper marks the initial exploration into the corpus's formulaicity, it is under the provisional hypothesis that such a correlation exists that we shall proceed. However, it is essential to consider this caveat for future linguistic and philological research.

Epigram entries in the DBBE include relevant metadata, such as the manuscript they are to be found in, the dating of that manuscript, their meter, etc. Of relevance to our discussion is the 'genre' that is attributed to them. The DBBE categorises the corpus into six genres 'based on the main actors that play a role in the communicative situation typical for book epigrams' (Ricceri et al., 2023). In this way, epigrams are divided into 'Text-', 'Author-', 'Scribe-', 'Reader-', 'Image-', and 'Patron-related', with epigrams often belonging to more than one genre. For the purpose of the present endeavour, this paper will focus on those 'Types' tagged as 'Scribe-related'.

The rationale behind this second choice can be summarised as follows. First, based on observations of former and current DBBE scholars, 'Scribe-related' epigrams, which are often referred to as 'metrical colophons', constitute the most standardised, 'formulaic' genre within the DBBE corpus (Bernard and Demoen, 2019). Furthermore, poetic colophons, constituting statements at the end of the book in verse, document information such as the scribe's name, the date of completion, and/or the place of writing. There is a practical need of providing a minimal set of information, which still provides room to poetic licence. Pre-constructed formulations are more likely to fulfil this specific yet standard need, with contextual information filling in 'slots' between standard expressions (see Example3.1). Lastly, metrical colophons constitute a well-studied genre due to their presence in various manuscript traditions, for example in Armenian manuscripts (e.g. Sanjian 1969; van Elverdinghe 2023). Therefore, conducting research on their Byzantine Greek counterparts will provide a solid foundation for future comparisons between these manuscript cultures.

4. Method

To achieve the objective of identifying and extracting formulaic sequences from our dataset, a multi-step methodology was devised. First, a conceptualisation of the DBBE formulae was undertaken, precisely defining the material sought after. Next, textual data was exported from the Scribe-related 'Types' subset of the database. This data then underwent pre-processing, before proceeding with n-gram extraction, wherein contiguous sequences of tokens were identified. These sequences were subsequently compiled into a dictionary, constituting a repository of recurring patterns. In the next stage, the dataset was refined, prioritising formulaic sequences and eliminating redundancy. Finally, manual evaluation was conducted. This entailed systematic comparison and documentation of formulaic material. This comprehensive methodology ensures a systematic and comprehensive approach to identifying and extracting formulaic sequences from the DBBE corpus. It addresses the complexities of the Byzantine book epigram corpus and navigates potential obstacles encountered during the analysis.

4.1. Conceptualisation: formulae in the DBBE

Before proceeding to the machine-assisted extraction of formulaic material from the DBBE corpus, it was crucial to define what constitutes a formula. This definition guided our search for relevant material and ensured consistency in our identification process. As acknowledged in scholarly discourse however, identifying formulae presents a challenge. This is due to the circularity of the task. As Wray (2008) puts it, 'you cannot reliably identify something unless you can define it. (...) In order to establish a definition, you have to have a reliable set of representative examples, and these must therefore have been identified first' (93).

Given that our corpus consists of written poems, distinct from the speech evidence typically used in modern linguistic research and the orality factor in Homeric studies, it was necessary to consider the following implications. In contemporary linguistic corpora, one of the functions of formulaic language is to reduce processing effort for speakers and listeners. This does not apply here as the DBBE corpus comprises written compositions. Instead, as suggested by Rutten and van der Wal (2012), the use of formulaic language in our corpus, and especially the Scribe-related subset, might serve as compensation for limited literacy and writing experience among scribes. Similarly, conclusions about 'speech community' practices or formulaic *speech* production cannot be drawn. Instead, it is possible to draw conclusions on the scribal 'text community' (Stenroos, 2018) and its repertoire, accepting the limitations of our historical corpus.

Considering these factors, it became essential to assess the applicability of common criteria used in linguistic research for identifying formulaic material:

1. Frequency: the frequency of a sequence remains a relevant criterion for the DBBE corpus.

2. Fixedness: while the degree of fixedness is applicable to our material, it is important to note that formulaic material may not always be fully-fixed (see Example 3.1). *Fully-fixed* formulae represent only one aspect of the spectrum of fixedness.

3. Intuition: intuitive judgements of native speakers regarding the formulaicity of sequences do not apply to historical corpora like ours. Although we could apply a notion of intuition,

the resulting conclusions would lack sufficient objectivity. Similarly, assessing compositionality or idiomaticity is challenging for non-native speakers, and depends on their degree of familiarity with the corpus.

4. While phonological aspects are not applicable, spacing and punctuation[6] can provide insights into 'chunking' patterns. However, this requires separate palaeographical analysis.

Based on these considerations, our working definition of formulaic sequences in the DBBE corpus is as follows: **recurring phrases that are integral to the repertoire of the scribal 'text community'**.

4.2. Data Export

First, textual data from the DBBE was extracted. At this preliminary stage, it was determined that exporting all DBBE 'Types' would suffice. As we already mentioned in Section 3, the 'Occurrences' are out of scope for this paper. From the 5,022 'Types', we compiled a collection of the 'Scribe-related Types'. For each of the remaining 1,849 epigrams we created a .txt file containing its text.

4.3. Pre-processing

Although the DBBE 'Types' are standardised, they present Byzantine Greek poems that employ an elaborate accentuation system, comprised of breathings indicated with a *spiritus asper* (rough breathing) or *lenis* (smooth breathing), and accents (acute, circumflex, grave), alongside nuanced punctuation conventions. While these linguistic features are relevant, they do not inherently influence the presence of formulaic material. The same applies to capital letters. However, they may pose challenges in character identification for computational analysis. For instance, distinguishing between Θεός and ϑεὸς, or Θεοῦ and Θ(εο)ῦ (expanded abbreviation), may be irrelevant for our analytical objectives. Therefore, pre-processing of the textual data was performed by removing accentuation, punctuation, and any non-essential formatting to ensure uniformity and enhance computational analysis capabilities.

(3) Χριστέ, ὁ ϑεὸς ἡμῶν χαροποιήσας
Christe, ho theos hèmōn charopoièsas
Christ, our god that causes joy.
DBBE Type 2380

(4) Ἀρχὴ καὶ τέλος ὁ Θεὸς ἡμῶν δόξα
Archè kai telos ho theos hèmōn doxa
The beginning and end, our God, our splendour
DBBE Type 4131

Without this pre-processing, the 3-gram ὁ ϑεὸς ἡμῶν from Example 3 would never match the 3-gram ὁ Θεὸς ἡμῶν from Example 4, even though they differ only in whether or not the thèta is capitalised. When counting n-grams, it is not desirable to miss out on correct matches due to these *irrelevant* linguistic features.

4.4. N-gram extraction

Subsequently, the pre-processed text underwent n-gram extraction. N-grams represent contiguous sequences of n tokens from a given sample of text, such as the 2-gram ὁ ϑεὸς in Example 3 or the 3-gram Ἀρχὴ καὶ τέλος in Example 4. We computed 2- to 12-grams for all 'Types' present in our dataset, providing frequency counts. The rationale behind the maximum n value of the n-grams was to capture the maximum number of words per verse present in our subset, which is 12. Based on our familiarity with the corpus, recurring patterns tend to occupy half a verse, a whole verse, or multiple complete verses. Thus, this approach allowed the identification of formulaic sequences present within verse limits.

At this stage, given the flexible syntax of the Greek language (van Emde Boas et al., 2019), two approaches to n-gram extraction were explored: one that takes into account the word order per verse and another that only considers the presence of words regardless of their order. In the former approach, variation in word order is captured and deemed significant, with the aim of capturing *fully-fixed* formulaic sequences. In contrast, the latter approach focuses on capturing a broader range of formulaic patterns (i.e. beyond the fully-fixed ones), thus assuming that word order is not a significant limiting factor in Byzantine Greek formulaic language as found in book epigrams.

Our familiarity with the corpus supports the latter perspective. The result of this step of the methodology consisted of lists of non-word-order-sensitive n-grams and their corresponding frequency.

4.5. Dictionary creation

Following the extraction of all n-grams, a dictionary comprising them was compiled to facilitate subsequent analysis. This dictionary functioned as a comprehensive repository of recurring sequences within the corpus, sorted from most to least frequently occurring. All combinations of 2 to 12 words

[6] Note that most texts are written in *scriptio continua* (i.e. without separating words using spaces) and the use of punctuation differs from our modern conventions.

that occur more than once in our corpus were included in the dictionary as potentially formulaic. It is noted that sequences with higher frequency counts are more likely to be formulaic. However, it is evident that a sequence of function words (i.e. articles, conjunctions, prepositions, pronouns), [7] cannot – for our purposes at least – feasibly be deemed formulaic, despite their frequent appearance in the corpus. In this paper, these are called *function-word sequences* and considered non-formulaic.

4.6. Last dataset refinement

In order to further refine the dataset and prioritise recurring sequences that are most likely to be formulaic, an additional automatic cleaning process was implemented. This process involved subtracting the frequency of (n+1)-grams that include an n-gram from the frequency of that n-gram. In essence, if an n-gram occurrs within an (n+1)-gram, the frequency count of the (n+1)-gram was deducted from the frequency count of the n-gram. For example, the 3-gram ωσπερ, ξενοι, χαιρουσιν occurs 13 times, while the 4-gram ωσπερ, ξενοι, χαιρουσιν, ιδειν occurs 12 times. Thus, the non-redundant frequency of this 3-gram is 1, as in all other 12 instances it occurs merely as a part of the 4-gram. This adjustment aims to reduce redundancy and enhance the distinction between phrases that are parts of formulaic sequences and complete formulaic units themselves.

4.7. Manual formulaicity evaluation

The final stage of analysis involved a comparative examination of the results based on the new n-gram values.

Interpreting the formulaicity results acquired from the previous step, required the following considerations:

1. *Function-word sequences* are not considered formulaic.

2. 2- and 3-grams containing at least one or two non-function words, respectively, are not considered formulaic; e.g., the bi-gram τε, και (both, and) occurring 119 times, την, βιβλον (the book) occurring 96 times, or η, βιβλος, αυτη (this book here) occurring 49 times. Exceptions are cases that represent a recognisable entity belonging to the Christian repertoire [8]. These can be prepositional phrases, like σὺν Θεῷ *syn Theō* (with (the help of) God), short supplications, like δίδου μοι *didou moi* (give me), or typical vocatives and exclamations, such as Χριστὲ μου *Christe mou* (my Christ) or δόξα Σοι *doxa Soi* (praise to You).

3. Special status is attributed to (>2)-grams containing articles or pronouns that do not modify terms within the (>2)-gram, prepositions without their modifier, or transitive verbs that render the (>2)-gram semantically incomplete without an object. These are considered potential *open-slot formulae*. The term *open-slot formula* will be henceforth used to refer to formulaic material that includes placeholders (X) to be filled based on the occasion (e.g. δόξα τῷ Θεῷ τῷ X (adj.) (*doksa tōi Theōi tōi X (adj.)*, praise to God, the X (adj.), also Example 3.1).

4. If the n-gram yields a positive frequency count, it is more likely to be the formula itself, with the (n+1)-gram being the formula accompanied by an element (e.g. an optional modifier) that frequently co-occurs, although the n-gram occurs more frequently as a standalone entity. For example, συν, θεω (with (the help of) God) occurs 56 times, and τελος, συν, θεω (the end, with (the help of) God) occurs 31 times. Based on the dataset refinement described above, the non-redundant frequency count of the bi-gram is 25. Thus, we can say that συν, θεω is the formulaic element, frequently but not exclusively paired with τελος in our subset.

5. If the n-gram count equals zero, both the n-gram and (n+1)-gram occur equally (i.e., only together). For example, the 4-gram ξενοι, χαιρουσιν, ιδειν, πατριδα occurs 10 times, and the 5-gram ωσπερ, ξενοι, χαιρουσιν, ιδειν, πατριδα also occurs 10 times. Thus, the 4-gram is deemed insignificant (non-redundant frequency count is set to 0) for further exploration as a stand-alone formula. In this case, the (n+1)-gram constitutes a formulaic sequence, and the n-gram is exclusively a part of it and, thus, not an independent entity.

6. If the n-gram presents a negative frequency count, this indicates that the n-gram occurs as part of one or more (n+1)-grams. This suggests that the n-gram is a common recurring pattern included in one or more formulaic (n+1)-grams, with other elements intervening. This is because the n-gram is captured in fixed order by default, but the subtraction procedure for eliminating redundancy considers the elements of the n- and (n+1)-grams in free order. In short, a negative frequency count reveals that the n-gram represents a formulaic element of lower hierarchy but is still related to the one represented by the (n+1)-gram. For example, the bi-gram χριστε, σωσον occurs twice, while

[7] e.g., καὶ οἱ *kai hoi* (and the), ἐν τῇ *en téi* (in the), or τῶν ἐμῶν *tōn emōn* (of mine)

[8] cf. Kuiper (2009) '+nlcu', Non-Linguistic Conditions of Use)

the 3-gram χριστε, μου, σωσον presents a frequency count of 15. In this case, the bi-gram's non-redundant count is set to -13, and it is considered a related, less frequent variant of the 3-gram. For our purposes, these n-grams will be referred to as *component-formulae*, reserving the term *formula* for the (n+1)-grams in this context.

7. Formulae can be maximum one verse long based on our Method (4.4). This means that *component-formulae* are shorter, while multi-verse entities are here considered a compilation of different formulae and called *patterns*.

Through a systematic comparison of n-grams across the corpus, recurring patterns suggestive of standalone formulaicity were isolated and documented. That resulted in a list of formulaic material within our DBBE subset, the frequency of which we acquired in step 3 and by n-gram extraction.

5. Results

Due to space constraints, this section will primarily discuss select yet representative findings.

The analysis will start from a well-known and established formula which recurs in Byzantine manuscripts, commonly known as the 'ὥσπερ ξένοι formula' (Example 1). Based on our last dataset refinement (see 4.6) and the criteria outlined above (in 4.7), it is confirmed that it constitutes a multi-verse pattern consisting of verse-long formulae.

The ὥσπερ ξένοι pattern's initial part/formula appears in two primary recurring forms (Table 2, A), with a total count of 14 occurrences in non-identical epigrams. At this level of analysis, there is no need to consider the difference between χαίρουσι (verb) and χαίροντες (participle), both stemming from the verb χαίρω ('to rejoice'). Similarly, the pattern's second part/formula (B, total frequency 6) exhibits minor spelling variations in the adverb (οὕτως / οὕτω). Although the formula is typically represented in scholarship by these two verses (as seen in Example 1), our results (Table 2) showed that this represents only half the truth in book epigrams, as A and B do occur consecutively (AB, see co-occ1), albeit with a frequency as low as 3 in our subset.

Formulae C and D represent structures mirroring formula A of the pattern, thus enriching the ὥσπερ ξένοι simile structure (e.g. ACDB). Among these, while D occurs often, the formulaic structures C, differentiated only by the use of different verbs (εὑρίσκω and ὁράω, respectively), collectively amount to double the frequency of D. This indicates that in cases of creative elaboration, the parallel structure/formula C is preferred. This preference aligns with Treu (1977), who suggests that C was the prototypical form of the formula, possibly dating back to the Greco-Roman period. Its earliest attestation, however, is to be found in the 9th century (*Palatinus gr.* 44), with formula A dominating only from the 10th century onward (*Parisinus gr.* 781).

Interestingly, our data shows that formula A does not exclusively replace formula C; rather, they frequently co-occur in several 'Types'. More specifically, A is combined with the most frequently occurring parallel structure (i.e. C) nine times, four of which are subsequently followed by the second part of the pattern (i.e. B).

The next example serves as another representative instance of a formula; through it, various aspects of the extracted formulaic material will be illustrated.

Firstly, akin to this example, most formulae are linked to and highlight the strong ritual aspect inherent in the corpus. As indicated in row 1 of Table 3, the 2-gram σὺν Θεῷ (*syn Theōi*, with (the help of) God) appears 56 times across different 'Types', solidifying its status as a prominent element in the Byzantine book epigram scribal repertoire. Given the deeply ingrained Christian context of Byzantine book production, it is unsurprising to observe such elements with high frequencies, with this particular 2-gram ranking as the 3rd most frequent formulaic sequence.

Moving on, row 2 presents what appears to be a half-verse formula. Comprising five syllables, it serves as an ideal sequence to occupy the first half-verse, with a caesura occurring after the 5th syllable, typical of the 12-syllable verse, the most frequently employed meter in the corpus.

Lastly, rows 3 and 4 offer examples of what we previously referred to as an *open-slot formula*. Half of the 3-grams (row 2) present a complement, as evidenced by the results of the formulaic 4-grams (row 3). Along with the absence of this pattern in the formulaic 5-grams, this suggests the presence of an empty slot, allowing for the inclusion of non-specific material, provided it constitutes a noun with or without modifiers (e.g., τέλος σὺν θεῷ τῆς θεολόγου βίβλου, 'the end, with (the help of) God, of the theological book'). Notably, the two instances of the 4-gram formula here are grammatically identical (τέλος σὺν Θεῷ ARTICLE). However, until a reliable part-of-speech tagger for Byzantine Greek is developed (Swaelens et al., 2023), researchers must manually identify this grammatical similarity.

6. Conclusion & Future Research

In this study, we have presented a methodology for identifying and extracting formulaic sequences from a subset of the Database of Byzantine Book Epigrams. The systematic application of this methodology offered a quantitative perspective of the prevalence and characteristics of formulaic language in

N-gram	F[9]	
ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα	10	A
ὥσπερ ξένοι χαίροντες ἰδεῖν πατρίδα	4	A
οὕτως καὶ οἱ γράφοντες ἰδεῖν βιβλίου τέλος	3	B
οὕτω καὶ οἱ γράφοντες ἰδεῖν βιβλίου τέλος	3	B
ἰδεῖν πατρίδα | οὕτως καὶ οἱ γράφοντες	3	co-occ1
καὶ οἱ θαλαττεύοντες εὑρεῖν λιμένα	10	C
καὶ οἱ θαλαττεύοντες ἰδεῖν λιμένα	3	C
πατρίδα | καὶ οἱ θαλαττεύοντες	9	co-occ2
λιμένα | οὕτως καὶ οἱ γράφοντες	4	co-occ3
καὶ οἱ στρατευόμενοι ἰδεῖν τὸ νῖκος	7	D

Table 2: The multi-verse ὥσπερ ξένοι pattern

N-gram	F
σὺν Θεῷ	56
τέλος σὺν Θεῷ*[10]	31
τέλος σὺν Θεῷ τῆς **X**	8
τέλος σὺν Θεῷ τοῦ **X**	7

Table 3: Ritual repertoire, half-verse and open-slot formulae

the Byzantine book epigram tradition. Our methodology provides a systematic framework for an initial step towards a comprehensive study of the formulaicity and creativity present in Byzantine book epigrams.

Nevertheless, our study is not without limitations. The methodology relies on computational analysis, which overlooks certain nuances or cultural contexts inherent in the corpus. Therefore, this paper proposes machine-assisted extraction of formulaic material as an initial step before engaging in linguistic and philological research. Additionally, the current focus on the DBBE subset limits the generalizability of our findings to the broader Byzantine epigram tradition, encompassing both book contexts and other mediums. Future research presents potential for further refinement of our methodology. Statistical analysis could aid in discerning patterns or collocations that form fixed sequences, distinguishing them from purely grammatical ones. Moreover, the creation of a reliable lemmatiser for Byzantine Greek could eliminate the distinctions between patterns that vary only in morphological details. Similarly, the development of a tool for addressing itacism would eliminate discrepancies arising from different orthographic representations. Lastly, we will also investigate automatic part-of-speech tagging for Byzantine Greek (Swaelens et al., 2023) to identify *open-slot formulae*.

7. Bibliographical References

References

Egbert Jan Bakker. 2005. *Pointing at the Past : From Formula to Performance in Homeric Poetics*. Cambridge : Harvard university. Centre for Hellenic studies.

Daphne Baratz. 2015. The repetitive structure in verse: A comparative study in homeric, south slavic, and ugaritic poetry. *Greek, Roman and Byzantine studies*, 55:1–24.

Klaas Bentein. 2023. A Typology of Variations in the Ancient Greek Epistolary Frame (I–III AD). In Georgios K. Giannakis, Panagiotis Filos, Emilio Crespo, and Jesús de la Villa, editors, *Classical Philology and Linguistics: Old Themes and New Perspectives*, pages 429–472. De Gruyter.

Floris Bernard and Kristoffel Demoen. 2019. Book epigrams. In Wolfram Hörandner, Andreas Rhoby, and Nikolaos Zagklas, editors, *A companion to Byzantine poetry*, volume 4 of *Brill's Companions to the Byzantine World*, pages 404–429. Brill.

Julie Boeten, Mark Janse, Klaas Bentein, and Ilse De Vos. 2021. *Byzantine Metre from the Margins : A Corpus-Based, Pragmatic Analysis of Medieval Book Epigrams*. Ph.D. thesis, Ghent.

Chiara Bozzone. 2010. New Perspectives on Formulaicity. In Stephanie Jamison, W., H.-G. Melchert, and Brent Vine, editors, *Proceedings of the 21st Annual UCLA Indo-European Conference*, pages 27–44. Hempen Verlag, Bremen.

Chiara Bozzone. 2014. *Constructions: A New Approach to Formularity, Discourse, and Syntax in Homer*. Ph.D. thesis, University of California, Los Angeles.

Andreas Buerki. 2016a. Automatic identification of formulaic sequences in (fairly) big data: practical

introduction to a procedure. *Advances in Identifying Formulaic Sequences:A Methodological Workshop*.

Andreas Buerki. 2016b. Formulaic sequences: A drop in the ocean of constructions or something more significant? In Ian McKenzie and Martin A. Kayman, editors, *Formulaicity and Creativity in Language and Literature*, pages 15–36. Taylor & Francis 2018.

Silvie Cinková, Pavel Pecina, Petr Podveský, and Pavel Schlesinger. 2006. Semi-automatic building of Swedish collocation lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Giovanni. Ciotti and Hang. Lin, editors. 2016. *Tracing Manuscripts in Time and Space Through Paratexts Perspectives from Paratexts*. De Gruyter.

Viviana Cárdenas Cortes. 2022. *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, chapter Corpus del español y lenguaje formulaico. Routledge.

Gilles-Maurice de Schryver. 2008. Why Does Africa Need Sinclair? *International Journal of Lexicography*, 21(3):267–291.

Kaja Dobrovoljc. 2020. Identifying dictionary-relevant formulaic sequences in written and spoken corpora. *International Journal of Lexicography*, 33(4):417–442.

John Miles Foley. 2007. "Reading" Homer through Oral Tradition. *College Literature*, 34(2):1–28.

Gérard Garitte. 1962. Sur une formule des colophons de manuscrits grecs. *Collectanea Vaticana in honorem Anselmi M*, 1:359–390.

Gerard Genette. 1987. *Seuils*. Seuil, Paris.

Adele E. Goldberg. 2006. *Constructions at Work the Nature of Generalization in Language*. Oxford : Oxford University Press.

A.E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. University of Chicago Press.

Elizabeth Jeffreys and Michael Jeffreys. 1983. The Style of Byzantine Popular Poetry : Recent Work. *Harvard Ukrainian Studies*, 7:309–343.

Michael Jeffreys and Elizabeth Jeffreys. 1986. The Oral Background of Byzantine Popular Poetry. *Oral Tradition*, 1(3):504–547.

Michael J. Jeffreys. 1973. Formulas in the Chronicle of the Morea. *Dumbarton Oaks Papers*, 27:163–195.

A. D. Kominis. 1966. *To Byzantinikon Hieron Epigramma kai hoi Epigrammatopoioi*. Athens.

Koenraad Kuiper. 1996. *Smooth Talkers : The Linguistic Performance of Auctioneers and Sportscasters*. Mahwah (N.J.) : Erlbaum.

Koenraad Kuiper. 2000. On the Linguistic Properties of Formulaic Speech. *Oral Tradition*, 15(3):279–305.

Koenraad Kuiper. 2009. *Formulaic Genres*. Houndmills.

M.D. Lauxtermann, Österreichische Akademie der Wissenschaften. Kommission für Byzantinistik, and Universität Wien. Institut für Byzantinistik und Neogräzistik. 2003. *Byzantine Poetry from Pisides to Geometres: Epigrams in context*. Byzantine Poetry from Pisides to Geometres: Texts and Contexts. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Maria Letizia Lazzarini. 1976. *Le formule delle dediche votive nella Grecia arcaica*. Roma : Accademia nazionale dei Lincei.

Martti Leiwo. 2005. *Ancient Greece at the Turn of the Millennium. Recent Work and Future Perspectives. Proceedings of the Athens Symposium, 18-20 May 2001*, chapter Substandard Greek. Remarks from Mons Claudianus. Publications of the Canadian Institute in Greece, No. 4.

Albert B. Lord. 1960. *The singer of tales*. Harvard University Press.

Claudine Moulin, Iryna Gurevych, Natalia Filatkina, and Richard Eckart de Castilho. 2015. *Historical Corpora: Challenges and Perspectives*, chapter Analyzing Formulaic Patterns in Historical Corpora. Narr Publishing House.

Delphine Nachtergaele. 2015. *The Formulaic Language of the Greek Private Papyrus Letters*. Ph.D. thesis, Ghent University.

Rachele Ricceri, Klaas Bentein, Floris Bernard, Antoon Bronselaer, Els De Paermentier, Pieterjan De Potter, Guy De Tré, Ilse De Vos, Maxime Deforche, Kristoffel Demoen, Els Lefever, Anne-Sophie Rouckhout, and Colin Swaelens. 2023. The database of byzantine book epigrams project: Principles, challenges, opportunities. *Journal of Data Mining & Digital Humanities*.

Gijsbert Rutten and Marijke van der Wal. 2012. Functions of epistolary formulae in Dutch letters from the seventeenth and eighteenth centuries. *Journal of historical pragmatics*, 13(2):173–201.

Avedis K. Sanjian, editor. 1969. *Colophons of Armenian Manuscripts 1301-1480: A Source of Middle Eastern History*. Cambridge.

J. Sinclair. 1991. *Corpus, Concordance, Collocation*. Describing English language. Oxford University Press.

Merja Stenroos. 2018. *Scribal Repertoires in Egypt from the New Kingdom to the Early Islamic Period*, chapter From Scribal Repertoire to Text Community: The Challenge of Variable Writing Systems. Oxford University Press.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023. Linguistic annotation of byzantine book epigrams. *Language Resources and Evaluation*.

Mariken Teeuwen and Irene van Renswoude, editors. 2018. *The annotated book in the early Middle Ages : practices of reading and writing*. Turnhout : Brepols.

Kurt Treu. 1977. *Der Schreiber am Ziel. Zu den Versen* Ὥσπερ ξένοι χαίρουσιν. . . *und ähnlichen*, pages 473–492. Berlin: Akademie-Verlag.

Emmanuel van Elverdinghe. 2023. *Armenia and Byzantium without Borders: Mobility, Interactions, Responses*, chapter The Hand That Once Wrote . . . ': The Journey of a Colophon Formula from Greek into Armenian. Brill.

Evert van Emde Boas, Albert Rijksbaron, Luuk Huitink, and Mathieu de Bakker. 2019. *The Cambridge Grammar of Classical Greek*. Cambridge University Press.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Alison Wray. 2008. *Formulaic language: Pushing the boundaries*. Oxford Applied Linguistics. Oxford University Press.