Advances in Archaeological Practice (2026), 1–17
doi:10.1017/aap.2026.10181

SAA
Society for American Archaeology

ARTICLE

Clarifying Stratigraphic Terminology in Paleolithic Archaeology Using Natural Language Processing

Jovan Galfi and João Cascalheira

Interdisciplinary Centre for Archaeology and Evolution of Human Behaviour (ICArEHB), Universidade do Algarve, Faro, Portugal
Corresponding author: Jovan Galfi; Email: jgalfi@ualg.pt

(Received 24 November 2025 UTC; revised 4 August 2026 UTC; accepted 7 August 2026 UTC)

Abstract

Archaeological context is a foundational concept, yet its definition remains ambiguous and inconsistent. Traditionally, lithostratigraphic frameworks rooted in geology have long dominated, but recent computational and spatial approaches increasingly redefine stratigraphic units. This shift raises a critical challenge: the lack of clear and consistent terminology for defining stratigraphic units. To explore this issue, we applied natural language processing (NLP) to a broad corpus of Paleolithic literature, focusing on the use of key terms such as "layer," "level," "occupation," and "horizon." Using SciBERT (Scientific BERT; BERT = Bidirectional Encoder Representations from Transformers) to generate contextual embeddings, we compared their semantic distributions with a control set of distinct terms (lithic, bone, flake, charcoal). The control set produced stable, separable clusters, whereas contextual terms showed extensive semantic overlap and instability, reflecting their interchangeable and inconsistent use. These results underscore that the problem is not only linguistic but theoretical: without clear terminological baselines, new methodological advances risk conflating analytical tools with interpretive outcomes. We argue that terminological precision is essential for comparability across sites and for cumulative research. NLP offers a replicable framework for auditing and clarifying terminology, contributing to greater rigor and interoperability in archaeological interpretation.

Resumen

El contexto arqueológico es un concepto fundamental, pero su definición sigue siendo ambigua y se aplica de manera inconsistente. Tradicionalmente, han predominado los marcos litoestratigráficos basados en la geología, aunque los enfoques computacionales y espaciales recientes están redefiniendo cada vez más las unidades estratigráficas. Este cambio plantea un problema crítico: la falta de una terminología clara y coherente para definir dichas unidades. Para examinar esta cuestión, aplicamos técnicas de procesamiento del lenguaje natural (PLN) a un amplio corpus de literatura sobre el Paleolítico, centrándonos en el uso de términos clave como layer, level, occupation y horizon. Utilizamos SciBERT para generar representaciones vectoriales contextuales y comparamos sus distribuciones semánticas con un conjunto de control formado por términos claramente diferenciados (lithic, bone, flake y charcoal). El conjunto de control produjo agrupaciones estables y separables, mientras que los términos contextuales mostraron un amplio solapamiento semántico e inestabilidad, lo que refleja su uso intercambiable e inconsistente. Estos resultados demuestran que el problema no es únicamente lingüístico, sino también teórico: sin bases terminológicas claras, los nuevos avances metodológicos corren el riesgo de confundir las herramientas analíticas con los resultados interpretativos. Sostenemos que la precisión terminológica es esencial para la comparación entre yacimientos y para el desarrollo de una investigación acumulativa. El PLN ofrece un marco reproducible para evaluar y aclarar la terminología, contribuyendo a una mayor rigurosidad e interoperabilidad en la interpretación arqueológica.

Keywords: context; data mining; machine learning; natural language processing; stratigraphy

Palabras clave: contexto; minería de datos; aprendizaje automático; procesamiento del lenguaje natural; estratigrafía

© The Author(s), 2026. Published by Cambridge University Press on behalf of Society for American Archaeology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

In a broad sense, context has always been part of archaeological research, but efforts to conceptualize it can be traced to the mid-nineteenth century (Trigger 2006). In the 1840s, Boucher de Perthes (1847) recognized the stratigraphic association of extinct fauna and stone tools and interpreted them as contemporaneous (Trigger 2006). Two decades later, Édouard Lartet (1861) used the superposition of associated assemblages to establish an early temporal organization of the Paleolithic. Although his periodization relied primarily on faunal remains, later researchers shifted the emphasis to associated stone-tool types, helping to establish periods still used today—such as the Acheulean, Mousterian, Aurignacian, Magdalenian, and Solutrean (Trigger 2006).

By the late nineteenth century, more careful consideration of spatial provenience emerged (Pitt-Rivers 1892). The shift stemmed less from methodological innovation than from a changed theoretical understanding of stratigraphic units. Artifacts began to be treated as meaningful assemblages reflecting the behavior of their makers (Boas 1932), and stratigraphic units became the primary units of comparison for understanding changes in human behavior (Lyman and O'Brien 1999). Seen this way, a stratigraphic unit is not only a methodological choice or terminological label; its use is grounded in theoretical assumptions.

Archaeological context in Paleolithic archaeology is commonly defined through stratigraphic units, reflecting the nature of the record, where features such as dwellings, pits, or graves are rarely preserved. For this reason, this article will refer to context as the smallest identifiable unit established during an excavation, defined by its stratigraphic position and sedimentary characteristics (Darvill 2003). In Paleolithic sites, these units are commonly equated with geological layers, separated by their color or contents, using postulates taken directly from geology (Goldberg and Macphail 2005; Lyell 1830–1833). Following rules such as superimposition, original horizontality, and lateral continuity of layers, researchers studying the human past create a robust and consistent context for the interpretation of past human behavior. Yet geological determinism has been repeatedly critiqued: because natural sedimentary processes are not equivalent to anthropogenic input, a fundamental dissonance in interpreting human activity is created (Schiffer 1987; Shahack-Gross 2017; Shott 1998). Although this problem is not unique to the Paleolithic, the scarce anthropogenic features found belonging to this period limit archaeological interpretation to these sedimentary boundaries for defining units of analysis. Geological determinism has long been the norm in Paleolithic research, but it risks conflation of units formed by natural depositional processes with discrete episodes of human activity. This is where terminology becomes important, because the labels used for these units can preserve or obscure the distinction between sedimentary formation and human behavior.

As far as we are aware, the pioneer of controlled documentation of Paleolithic stratigraphy was Dorothy Garrod in the 1920s through excavation in 5–10 cm spits within geological layers (Garrod 1937). Further advances, from profile recording and the Harris matrix (Harris 1979) to the adoption of total stations in the 1980s, created the infrastructure for modern computational analysis of stratigraphic and spatial data (McPherron et al. 2005). Building on this infrastructure, computational methods have emerged as powerful tools for interpreting spatial distributions and are increasingly used to redefine stratigraphic units after excavation (Arteaga-Brieba et al. 2023; Canals et al. 2003; Discamps et al. 2023; Falcucci et al. 2024; Mielgo et al. 2024; Sossa-Ríos et al. 2024).

Here, we will use post-excavation stratigraphy (PES), following Discamps et alia (2023), as an umbrella label for these approaches. PES approaches respond to problems that Paleolithic archaeology has long recognized—including time-averaging, palimpsest formation, and the effects of different site-formation processes (Bailey 2007)—by stepping back from geology and integrating new or existing datasets into the interpretation of sequences. In doing so, PES approaches allow Paleolithic archaeologists to establish smaller contextual units and to observe human behavior at finer temporal scales. Although still in their early stages, PES techniques have shown promise in addressing the limitations of lithostratigraphy. At the same time, they also add complexity; it remains unclear how these new stratigraphic units fit within established frameworks for defining stratigraphy, and how they relate to the lithostratigraphic units they are meant to refine. This uncertainty leads us to a fundamental question: What do we mean when we say "context?" Despite its foundational role, contextual units in Paleolithic

archaeology often rest on inherited, inconsistently applied criteria. As new technologies and analytical frameworks emerge, we must reexamine both traditional definitions and modern alternatives, beginning with a stable and explicit terminological baseline capable of signaling, managing, and comparing different concepts.

Stratigraphic Semantics: Considering Context in Paleolithic Archaeology

Formation processes theory (Schiffer 1972, 1987) has long emphasized that the archaeological record results from the interaction of cultural and natural processes, and that the distinction between systemic context and archaeological context is crucial when interpreting stratigraphy. From this perspective, lithostratigraphic units represent only one part of a more complex formation history that also includes postdepositional and cultural transformations.

In this light, researchers are making efforts to integrate methods of establishing context through the application of different clustering techniques (Clark 2017), lithic or faunal refitting (Deschamps and Zilhão 2018; Sossa-Ríos et al. 2024), machine learning (Arteaga-Brieba et al. 2023), and the integration of taxonomy/typology/taphonomy along vertical sequence of a site (Discamps et al. 2023; McPherron et al. 2005; Mielgo et al. 2024). Although these techniques retain the spatial logic of horizontal continuity and superposition, they are also marked by two key differences: they are defined from the spatial relationships among archaeological remains rather than sedimentary properties, and they are typically established or finalized after excavation is complete. However, such advances expose a persistent problem: the lack of clear terminological boundaries. Without explicit definitions that specify both the analytical purpose of a unit and the methodological criteria by which it is established, concepts risk being conflated or applied inconsistently. This undermines reproducibility and can lead to conflicting interpretations (Bicho and Cascalheira 2020).

An illustrative case is the effort to associate Nubian technology and Neanderthals at Shukbah cave (Blinkhorn et al. 2021). Blinkhorn and colleagues use varying terminology related to Shukbah's sequence, which includes "layer," "occupation," and "archaeological horizon" interchangeably. In contrast, the original excavator of the site, Dorothy Garrod, exclusively uses "layers" as the main contextual units, describing them as units of different geological attributes (Garrod 1942). In their reply, Hallinan and colleagues (2022) argue that Shukbah's stratigraphic units represent time-averaged deposits and as such cannot support a secure association between the Neanderthal remains and the Nubian artifacts. The exchange shows that questioning the chronological resolution of a deposit does not necessarily challenge the broader assumptions used to define archaeological context. Addressing this issue requires recognizing that methodological innovation alone cannot resolve ambiguity unless it is grounded in a coherent and explicit theoretical framework.

What's in a Name?

Although definitions of stratigraphic and contextual units exist (Darvill 2003; Gasche and Tunca 1983; Harris 1979), their application often depends on the author's background (Gilbert and Goldberg 2017). This points to a deeper issue in archaeological interpretation, because terminology should convey both the analytical purpose of a unit and the methodological criteria used to establish it. Instead, definitions are often inferred indirectly from archaeological dictionaries (Darvill 2003; Kipfer 2021) or excavation methodology manuals (Harris 1979; Renfrew and Bahn 2004).

As a result, researchers frequently use different terms when describing the same stratigraphic unit (SU), or the same term for units established through very different methods. Terms such as "level," "layer," or "horizon" may appear superficially interchangeable, yet the methods used to define them can vary substantially (Table 1), particularly with the rise of PES approaches. More complexity arises when these are conflated with interpretative terms such as "occupation," "episode," or "event." A "layer" identifies a bounded deposit or spatial distribution, whereas an "occupation" implies human presence and activity within that unit. "Episode" suggests a more discrete phase of activity, whereas "event" implies an even narrower and more closely contemporaneous occurrence; related concepts such as living floors carry similar fine-scale behavioral implications (Malinsky-Buller et al. 2011; Vaquero 2008; Vaquero et al. 2012). For example, there is often no explicit step in determining when a layer bearing Magdalenian

Table 1. Definitions, Analytical Bases, and Common Inferential Issues for Key Stratigraphic and Contextual Terms in Paleolithic Archaeology.

Term	Defined as	Based on	Issues	References
Layer	Geological unit based on physical traits of sediments	Observations in the field; geoarchaeological analysis	Often treated as an occupation proxy; boundaries can be coarse/time-averaged; sometimes "layer" is used as a generic synonym for "level" or "horizon"	Gasche and Tunca 1983; Goldberg and Macphail 2005; Harris 1979
Level	Excavation unit or slice of arbitrary thickness	Methodological choices	Defined as a nongeological excavation unit but often treated as a stratigraphic entity in writing; due to its mechanical nature, it can cut across lithostratigraphic boundaries	Gasche and Tunca 1983; Harris 1979
Horizon	A thin horizontally traceable feature or zone; often a correlation unit rather than a deposit unit	Observations in the field; interpretation of materials	It can be defined by both geological and archaeological entities in the sequence	Gasche and Tunca 1983
Occupation	Interpretive unit implying human presence/activity; not a sedimentary unit by default	Interpretation of materials	A stratigraphic unit is often equated to occupation; implies event/episode but usually time-averaged; often inferred with weak operational criteria	Bailey 2007; Bicho and Cascalheira 2020; Discamps et al. 2023; Machado et al. 2013

tools becomes a Magdalenian occupation, despite the very different inferential weight these terms carry. The former may only designate the stratigraphic unit in which Magdalenian material was recovered, whereas the latter implies that the material represents a meaningful phase or episode of human activity. In practice, they are often used interchangeably to denote human presence (Bailey 2007). When this inferential transition is left unstated, sedimentary, excavational, analytical, and behavioral units may be treated as equivalent despite being defined through different forms of evidence. This matters because archaeological analysis routinely operates across multiple unit types (e.g., layers, levels, assemblages, phases) that organize time and structure subsequent interpretation (Bailey 2007; Discamps et al. 2023). Formation-process theory has long demonstrated that archaeological deposits are not direct reflections of past behavior (Schiffer 1972, 1987); consequently, different kinds of units support different levels of behavioral or temporal inference.

Efforts to impose greater terminological precision are not new. In the 1970s and 1980s, Fedele (1976, 1984) proposed the concept of the "elemental sediment unit" (ESU), defined as the smallest geologically recognizable entity. Around the same time, Gasche and Tunca (1983), following a series of workshops, developed the *Guide to Archaeostratigraphic Classification and Terminology*. Their framework defined the lithologic unit as the base stratigraphic unit: a three-dimensional homogeneous entity composed of a dominant lithologic type or a consistent combination of several types. Both initiatives highlighted a central problem: semantic inconsistency, in which identical units were given different names, or distinct units were referred to with the same term.

Addressing such inconsistencies is especially important today, given that nonlithological approaches to defining stratigraphy have become increasingly prominent. Incorporating new types of SUs into existing interpretive frameworks without terminological standardization risks conflating methodological tools with interpretive outcomes. Establishing consistent baseline vocabulary would not only align Paleolithic archaeology with calls for greater standardization in geoarchaeology (Stein and Holliday

2020) but would also mirror successful efforts in other disciplines. Controlled vocabularies in ecology, linguistics, and the biomedical sciences, for example, have been crucial for ensuring interoperability and cumulative research (e.g., American Psychological Association 2001; Ashburner et al. 2000; Farrar and Langendoen 2003). Despite recognition of this need, a standard practice for defining and naming SUs in Paleolithic archaeology has yet to be implemented.

Terminological Instability in Archaeological Contexts: A Case for NLP Integration

The persistent problem of vague and inconsistent terminology in archaeology requires systematic analysis across a wide range of subfields and research traditions. Conducting such a review manually would be prohibitively time consuming, given the breadth of the archaeological corpus. Natural language processing (NLP) is a field concerned with the computational analysis of human language. Drawing on linguistics, computer science, and artificial intelligence, it enables large-scale analysis of terminology across diverse publications (Devlin et al. 2019; Martín-Rodilla 2023).

Large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-Trained Transformer) are well-established tools that encode language and represent it as measurable vectors (Brown et al. 2020; Devlin et al. 2019; Mikolov et al. 2013). Although generative models have recently become integral to producing scientific text, other capabilities of NLP (e.g., semantic similarity detection, clustering, terminology mapping) remain underutilized in archaeology (see Brandsen 2023; Martín-Rodilla 2023).

Because data mining and NLP can rapidly process vast corpora, detecting language patterns, sentiment, or named entities, they are increasingly applied in the social sciences (Franzosi et al. 2022; Macanovic 2022). Recent studies have used LLMs to investigate gender and ethnic stereotypes (Garg et al. 2018), cultural markers of class (Kozlowski et al. 2019), and representations of intangible cultural heritage (Ning and Chen 2023). In archaeology, ArchaeoBERT has demonstrated promise, demonstrating effectiveness in tasks such as named entity recognition (NER) and fill-mask completion. This model was fine-tuned on archaeological texts in English, Dutch, and German (Brandsen 2024; Brandsen et al. 2022).

Our analysis focuses on the classification of predefined terms related to archaeological context. Each term was observed within a defined window of neighboring words, converted into a high-dimensional vector, and then visualized using dimensionality reduction techniques such as t-SNE (Van der Maaten and Hinton 2008), Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2020), and principal component analysis (PCA). By systematically mining archaeological literature in this way, NLP can help standardize terminology, highlight inconsistencies, and generate tangible representations of implied or assumed meanings.

Materials and Methods

Data Mining and Preprocessing

To compile a dataset suitable for semantic analysis of archaeological texts, we implemented a Python workflow that automated data collection, cleaning, and formatting of publications for NLP. We used the Scholarly Python library to interact programmatically with Google Scholar, creating an Excel database that allowed for systematic screening of the queried articles. Because bulk downloading of articles can lead to IP address bans from Google Scholar, each article was manually downloaded using the URLs provided in the metadata. Although time consuming, this approach preserved the integrity of the data collection process and allowed for further control and screening of the sampled articles. The data collection was performed in July 2024, which defines the temporal boundary of the dataset.

Using the search string "paleolithic+occupation," we retrieved 960 records from Google Scholar. Because query-based searches introduce topical bias, this focus was deliberate: the aim was to examine terminology in occupation-related Paleolithic contexts. For each record, we documented author, year, title, and URL. Study selection followed PRISMA-style reporting principles (Page et al. 2021; Figure 1). Records were screened by title and abstract, and subsequently, by full text. Articles were included if they were peer-reviewed and addressed Paleolithic archaeology. No temporal restrictions were applied

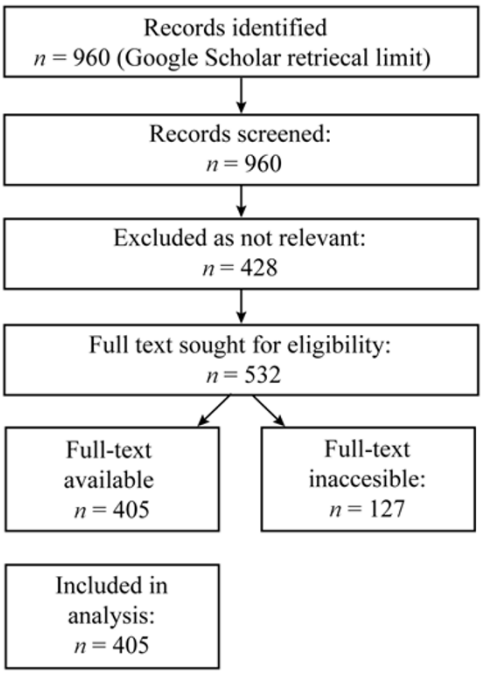


Figure 1. PRISMA-style flowchart outlining the corpus selection and screening process. Adapted from the PRISMA 2020 flow diagram (Page et al. 2021).

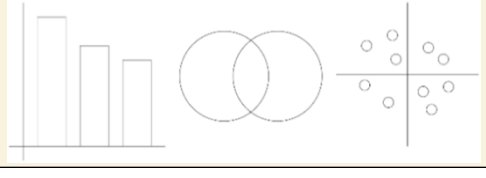
so as to capture long-term patterns of terminological usage. Purely geological or nonarchaeological publications were excluded. Although the search query centered on "occupation," eligibility remained broad within Paleolithic research (including site reports, syntheses, and methodological studies) to capture varied uses of stratigraphic terminology.

The collected articles were processed according to a workflow adapted from Hickman et alia (2022) and Brandsen (2023) for extracting text from scientific articles (Table 2). The library PyMuPDF was used to extract the text from articles and prepare it for further processing. As an initial step, the first page and references were removed. Stopwords and words shorter than three characters were removed, and the remaining words were tokenized—that is, transformed into individual units using the Natural Language Toolkit (NLTK) library (Bird et al. 2009). Each token was then lemmatized (e.g., "running" to "run"). We established the LBFC dataset (lithic, bone, flake, charcoal) under the assumption that these four terms are semantically distinct material-related terms. Although "flake" and "lithic" are related terms, including them was done intentionally to test whether the NLP pipeline can detect more nuanced distinctions within closely related terminology. For the main analysis, we generated the OLLH dataset (occupation, layer, level, horizon), based on the terms identified above as central to defining archaeological context.

NLP and Semantic Analysis

Our analysis is based on SciBERT, a version of BERT trained on a large corpus of open-access scientific articles (Beltagy et al. 2019). We opted not to use ArchaeoBERT (Brandsen 2024) given its preprint status and the unavailability of the fine-tuning corpus, which constrains reproducibility. SciBERT is limited to sequences of up to 512 tokens, constraining the analysis of scientific texts to abstracts (Danilov et al. 2021). To mitigate this, we set up smaller windows of tokens containing the LBFC and OLLH terms, testing the model on 10, 20, 30, and 40 token windows. The embeddings of all tokens in that window from the model's final hidden layer were averaged to produce a single 768-dimensional vector per occurrence. We also tested windows of 10, 20, 30, and 40 tokens for each term, observing similar patterns across settings (Supplementary Figures 1 and 2). The chosen range aligns with prior findings

Table 2. Workflow for Text Extraction, Semantic Analysis, and Visualization of the Results.

Action	Example	Resources Used
1. Access an article: • Create database • Download articles • Remove first page • Remove references	Sample text from Discamps et al. 2023: "definition of the most precise stratigraphic layers possible during fieldwork, usable criteria are obviously limited to what can be observed in situ in the excavation window. In practice, the recovered archaeological material assigned to these field layers is subsequently grouped into 'assemblages,' which are then stored in museums following these units (i.e., they contain homogeneous material), to reflect distinct occupation phase"	Google Scholar Scholarly PyMuPDF Regex
2. Process: • Lowercase • Stopword removal • Citation removal • Lemmatization	Article = ["definition," "precise," "stratigraphic," "layer," "fieldwork," "usable," "criterion," "obviously," "limit," "observe," "situ," "excavation," "window," "practice," "recover," "archaeological," "material," "assign," "field," "layer," "subsequently," "group," "assemblage," "store," "museum," "follow," "unit," "reflect," "occupation," "phase"]	Natural Language Toolkit (NLTK) Regex
3. Query: • Locate terms • Tokenize terms (within a 20-word window)	Layer = ["limit," "observe," "situ," "excavation," "window," "practice," "recover," "archaeological," "material," "assign," "field," "layer," "subsequently," "group," "assemblage," "store," "museum," "follow," "unit," "reflect," "occupation," "phase"]	Transformers
4. SciBERT: • Vectorize the window • Each term represented by a 768-dimension vector	Layer = [0.51, 0.12, −0.34, 0.29, 0.04, 0.18, −0.22, 0.06, 0.45, −0.18, 0.22, 0.37, −0.09, 0.13, −0.12, 0.09, 0.67, 0.31, −0.41, 0.58, 0.22, 0.25, . . . , −0.25, 0.27, −0.12, 0.15, −0.08, 0.04, 0.28 −0.16, 0.23, 0.42, −0.07, 0.10, −0.15, 0.07, 0.56, 0.07, −0.18, 0.31, 0.11, 0.20, −0.19, 0.05]	Transformers PyTorch
5. Data analysis and visualization: • Venn diagram • UMAP • Silhouette score • k -NN		Matplotlib scikit-learn

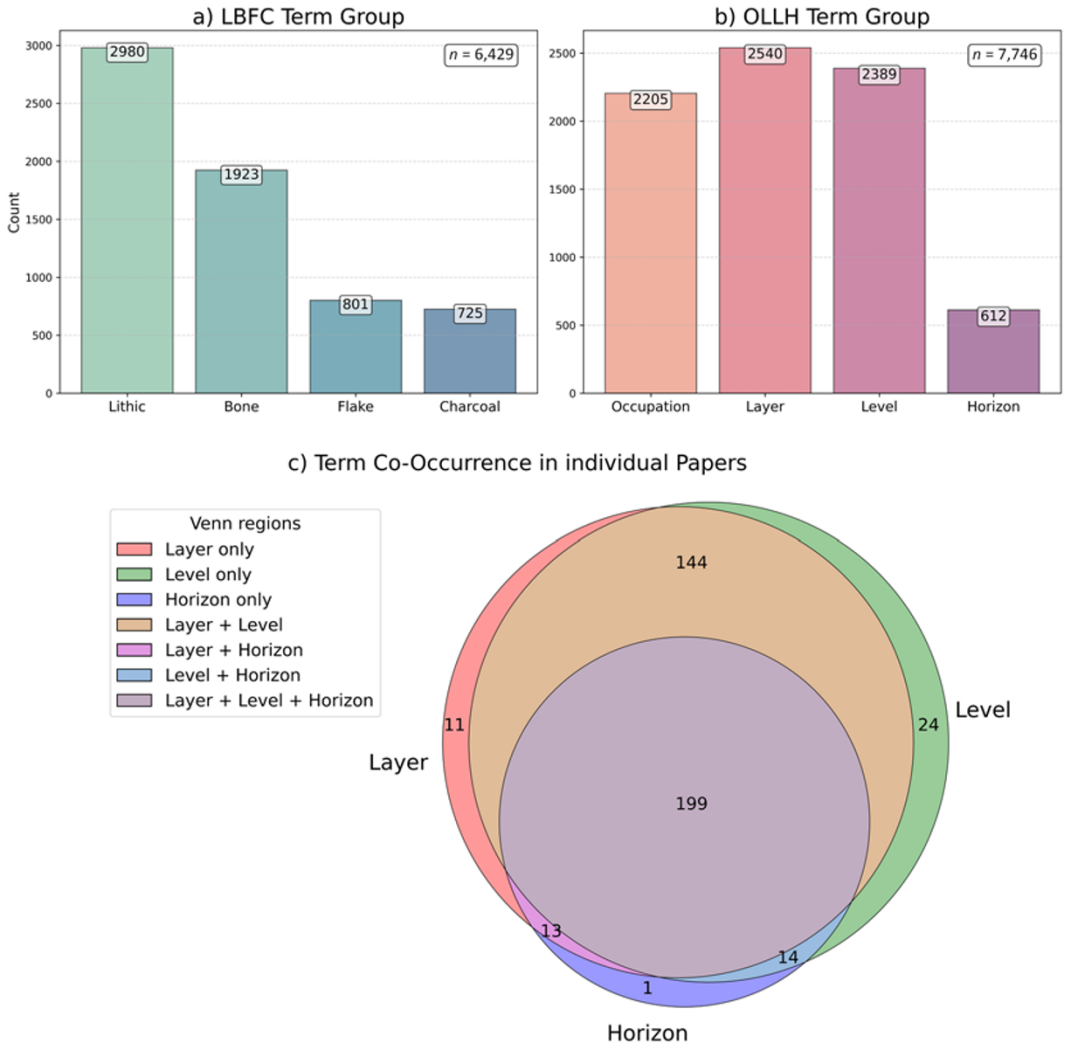


Figure 2. Results of data mining. A total of 14,175 individual terms were extracted from 405 scientific articles: (a) the test -LBFC group, (b) terms used for semantic analysis -OLLH group, (c) Venn diagram of the co-occurrence of the key contextual terms across individual articles (note that the term "occupation" is not included). Figure prepared by Jovan Galfi.

that context windows between 10 and 50 tokens yield the best clustering performance, balancing local term meaning and overall sentence semantics (Subakti et al. 2022).

The distribution of the semantic patterns of different terms was displayed using Uniform Manifold Approximation and Projection (UMAP). It is a widely used dimensionality-reduction technique given that it balances scalability to large datasets with relatively low computational cost (McInnes et al. 2020). It preserves relationships from high-dimensional space by identifying the nearest neighbors and maintaining them in reduced space, managed by two parameters: predetermined number of neighboring points and minimum distance allowed between the points. The minimum distance is defined as a proportion of the normalized distance scale established by UMAP, and we used the recommended value of 0.1, which best preserves both local and global relationships (McInnes et al. 2020). To evaluate robustness, we tested neighborhood sizes of 15, 30, 50, and 75, capturing both fine-grained local structure (lower values) and broader global patterns (higher values). To test the quality of the UMAP clusters, we have relied on silhouette score (Cunningham and Delany 2022; Rousseeuw 1987), proven reliable in testing whether the structures created by UMAP reflect meaningful patterns or artifacts of

dimensionality reduction (Dalmia and Sia 2021; Saha 2023). K-Nearest Neighbor (k-NN) evaluation used stratified 10-fold cross-validation (Kohavi 1995) implemented with scikit-learn (StratifiedKFold), with KNeighborsClassifier (k = 5, cosine) and metrics reported as per-term F1, Macro-F1 (mean ± s.d.) and overall Accuracy (Pedregosa et al. 2011); permutation p-values followed a label-permutation procedure (Ojala and Garriga 2010).

Results

Data Mining

The data collection process produced an initial set of 960 records from Google Scholar, which were compiled into a structured datasheet to guide manual article retrieval. After accounting for access limitations and relevance screening, the final dataset consisted of 405 articles. From these, a total of 14,175 individual term occurrences were extracted, divided between the LBFC group ("lithic," "bone," "flake," "charcoal") with 6,429 terms and the OLLH group ("occupation," "layer," "level," "horizon") with 7,746 terms (Figure 2).

Among these, "lithic" is the most frequent (n = 2,980), followed by "layer" (n = 2,540), "level" (n = 2,389), "occupation" (n = 2,205), "bone" (n = 1,923), "flake" (n = 801), "charcoal" (n = 725), and "horizon" (n = 612; Figure 2a–b). Visualizing co-occurrence revealed how often these terms appear within the same article (Figure 2c). Although this does not necessarily mean that they are used interchangeably, it indicates the frequency with which they are invoked together. Of particular note, 188 articles (46.5%) include three of the terms ("layer," "level," and "horizon"). A further 140 (33.6%) contained both "layer" and "level," which are used together in 328 (81%) articles. The term "occupation" was not visualized with the Venn diagram given that all of the articles contain it.

The Venn diagram highlights how rarely these terms appear in isolation: only 36 articles (8.8%) employed them individually (Figure 2c). Although it is the least frequent term overall (4.3%), "horizon" still appeared in more than half of the articles (n = 215), suggesting a distinct pattern of use compared to the other terms.

NLP and UMAP Analysis

The results of our analysis display two distinct patterns between the LBFC and OLLH datasets. The two datasets are shown to be distinct in both UMAP (Figure 4, Supplementary Figure 1), as well as k-NN, which displays high separability between datasets, with an accuracy of 0.945 ± 0.006 and a macro-F1 of 0.945 ± 0.006 (Figure 4).

LBFC Dataset

Testing the NLP pipeline on the LBFC dataset demonstrated the expected outcome: clear semantic separation among terms. Across all *n_neighbors* settings, the clusters remained stable in structure, with only their positions shifting (Figure 4a, c, e, g). Contextualized embedding clusters for "lithic," "flake," and "charcoal" remain proximate (Figure 4a), but their separation remains stable with increasing *n_neighbors*. Contextualized embeddings were consistently distinct across all settings, forming a well-defined cluster throughout. The silhouette scores remain between 0.48 and 0.49, showing moderate but meaningful separation of terms. This is confirmed by k-NN, displaying a very high accuracy (Figure 3; Accuracy = 0.952 ± 0.009; M-F1 = 0.937 ± 0.012)

Lithic contextualized embeddings produced two separate clusters, which become more distant with the increase of *n_neighbors* (Figure 4a, c, e, g). Clusters of "lithic," "flake," and "charcoal" remain proximate while maintaining separation. They start to diffuse as the *n_neighbors* value increases, implying that their separation is stronger at a more global scale. The rest of the term embeddings remain separated and stable throughout different *n_neighbors* values.

Although the clusters remained stable overall, limited intermixing was observed, such as small numbers of "bone" and "charcoal" contextualized embeddings appearing within the larger lithic cluster. This behavior is consistent with the design of the embeddings, which capture semantic context rather than strict lexical similarity, and it reflects the heterogeneous character of the source literature.

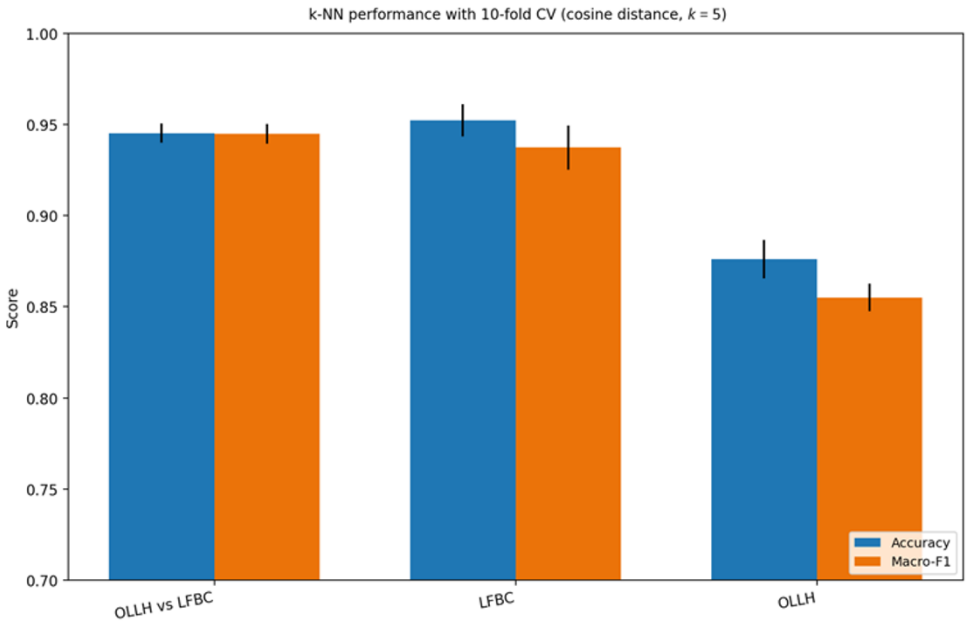


Figure 3. Cross-validated k-NN results.

OLLH Dataset

In contrast, the OLLH dataset produced a markedly different pattern. Unlike the LBFC dataset, the embeddings for "occupation," "layer," "level," and "horizon" failed to form distinct clusters, instead appearing widely dispersed with significant overlap (Figure 4b, d, f, h), with the layer cluster in particular collapsing into the other embeddings. Across all *n_neighbors* values, the results displayed extensive intermixing with no stable separation of terms, reflected in a much lower silhouette score (0.06–0.08). Although there are smaller separate clusters present, they are made up of intermixed contextualized terms. Although the value for the k-NN test remains fairly high (Accuracy 0.876 ± 0.011; Macro-F1 0.855 ± 0.008; $p = 0.005$), their UMAP overlap is too extensive. The clusters maintain the same intermixed structure throughout, showing that the intermixed structure of the OLLH embeddings is inherent and independent of projection settings.

Summary

Comparing the two datasets highlights a sharp contrast: whereas LBFC embeddings reveal stable and distinct clusters, OLLH embeddings show extensive overlap, suggesting that these terms are often used interchangeably in archaeological literature. Although the results of k-NN display high accuracy for the OLLH dataset (Figure 3), it is still lower relative to the LBFC dataset. This high accuracy is not reflected in the UMAP projections, which maintain low silhouette scores and show extensive overlap. This apparent contrast reflects the difference between local and global structure in the data: the k-NN classifier captures local distinctions in high-dimensional space, whereas the low silhouette scores and UMAP projections indicate that these distinctions do not translate into well-separated clusters at a global level.

This interpretation is reinforced by the co-occurrence data, which shows that terms in the OLLH dataset are rarely employed independently. The absence of terminological structure in the embeddings therefore directly mirrors the semantic ambiguity documented in Section 1, underscoring the need for clearer definitions and standardization in the use of contextual terminology.

Discussion

The results of this study demonstrate both the potential and the limitations of applying NLP to archaeological terminology. Using SciBERT within a structured workflow, we were able to establish a replicable

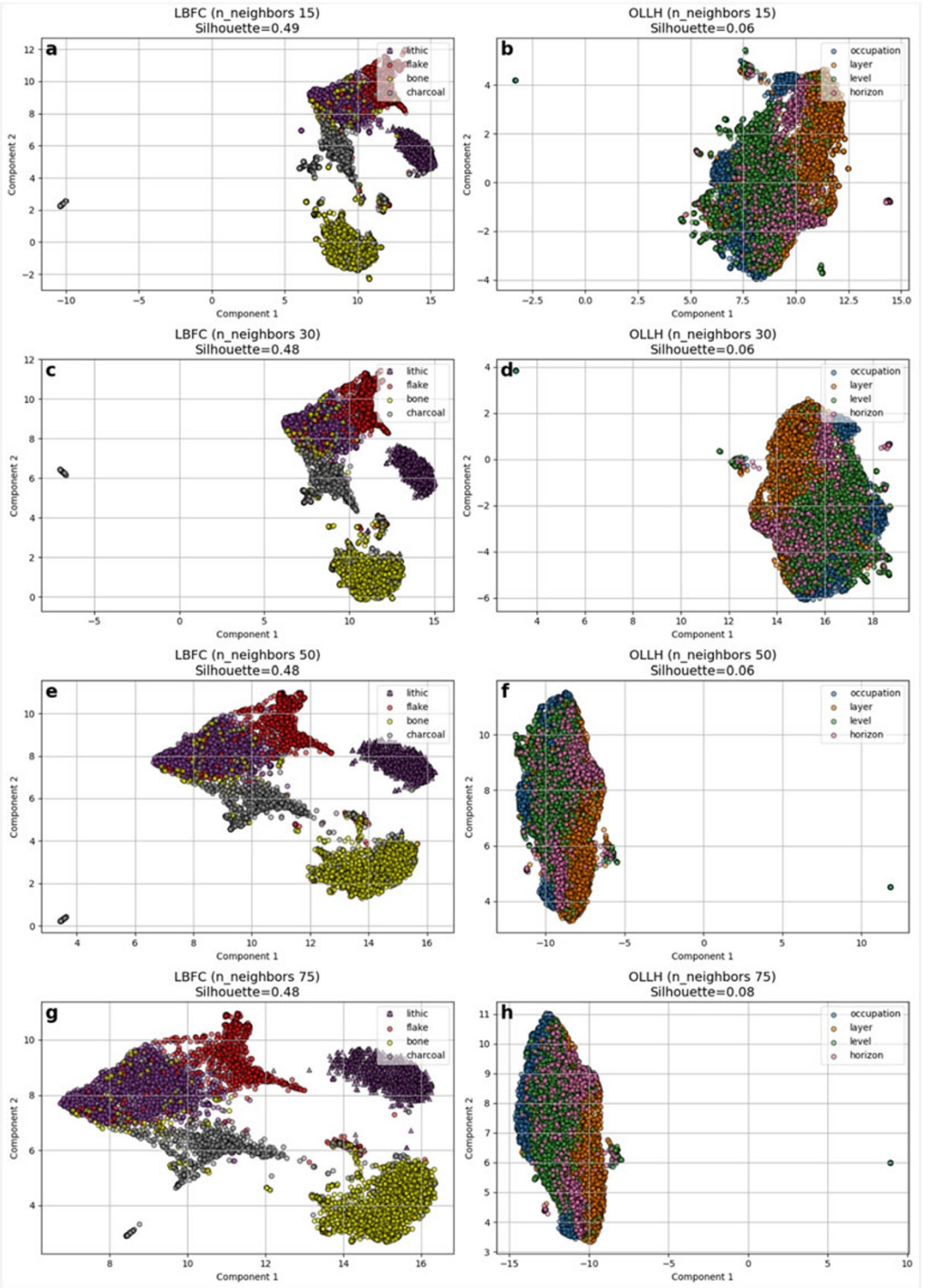


Figure 4. UMAP structure across different neighborhood sizes.

framework for analyzing scientific texts and their use of contextual terms. The LBFC dataset performed as expected: embeddings separate into stable clusters across neighborhood values, with additional nuance captured in the bifurcation of "lithic" embedding. In contrast, the OLLH dataset revealed semantic instability, with overlapping embeddings reflecting that they are applied in closely related contexts. These results mirror the interpretative challenges outlined above, where inconsistent terminology undermines the capacity to establish clear archaeological contexts. It is important to note that semantic overlap in embedding space does not by itself demonstrate interchangeability. Terms such as "layer" and "occupation" frequently co-occur in hierarchical constructions (e.g., "the occupation identified in level B") in which they operate at different analytical scales rather than as synonyms, and our approach cannot distinguish such structured co-occurrence from genuinely synonymous usage. Our interpretation of the OLLH collapse as terminological instability therefore rests not on the embeddings alone but on their convergence with the qualitative evidence discussed at the beginning of the article, where the same terms are demonstrably applied to units established through different criteria.

Importantly, our results provide quantitative evidence of semantic instability in archaeological terminology. Whereas earlier critiques (e.g., Schiffer 1987; Shott 1998) highlighted these problems qualitatively, the NLP-based analysis empirically demonstrates the scale of overlap among terms such as "layer," "level," "occupation," and "horizon." This strengthens the argument that terminological ambiguity is not anecdotal but systemic, and it can now be visualized and measured.

The persistence of synonymy and polysemy in archaeological discourse is particularly evident in the UMAP projections (Figure 4). Proximity and cluster overlap are consistent with synonymy, whereas polysemy is suggested by dispersed points. This reflects a broader disciplinary pattern: stratigraphic units are often described in everyday language, relying on assumed meanings rather than explicit definitions. Such ambiguity may have been manageable when context was defined primarily by geological principles, but with the advent of nongeological approaches to stratigraphic units, the risks of conflation are higher. Terminology must therefore remain purposeful, both in local excavation contexts and in comparative, cross-site analyses. When unit vocabulary is unstable, the consequences are concrete: cross-site correlations may equate units that are not analytically equivalent; behavioral attributions may rest on units that cannot support them, as at Shukbah, where a species–technology association stands or falls on what "layer" designates; and aggregation platforms inherit incommensurable units, undermining the comparative queries they exist to enable. Terminological instability is therefore systemic rather than anecdotal, with direct consequences for how contexts are defined, compared, and interpreted.

This concern connects to long-standing theoretical debates. Concepts such as time-averaging, palimpsests, and formation processes (Bailey 2007; Binford 1981; Lucas 2012; Schiffer 1972, 1987) have long framed archaeological interpretation, and the coarse temporal resolution of the record itself constrains the inferences that analytical units can support (Perreault 2019). However, archaeology has been slow to fully integrate its own material datasets into stratigraphic reasoning.

For much of its history, Paleolithic archaeology has relied on geoarchaeology as the primary mediator of context (Shahack-Gross 2017). Stratigraphy has been the backbone of cultural succession models since the inception of the field, with cultures defined as fossil-like entities aligned to geological layers. The introduction of PES approaches, however, means that stratigraphic units are no longer restricted to lithological criteria; they are multidimensional entities constructed from spatial, typological, and taphonomic information. While promising, this transition carries risks: without theoretical integration, PES can collapse into self-reinforcing interpretive loops, where behavioral inferences are used to justify the very contexts from which behaviors are inferred (Discamps et al. 2023; Mielgo et al. 2024; Sossa-Ríos et al. 2024). If Paleolithic archaeology is to strengthen its claim to scientific rigor, such inconsistencies must be addressed (Marwick 2025). The central issue is not whether PES and similar approaches should be adopted—they clearly expand the methodological repertoire—but whether they should be described using the same terminological vocabulary as lithostratigraphic units. To do so risks conflating methodological procedures with interpretive outcomes. Establishing clear terminological boundaries is therefore essential to avoid undermining interpretability and comparability across sites. In this sense, our results provide quantitative support for concerns that archaeological theory has articulated qualitatively for decades (Bailey 2007; Lucas 2012; Malinsky-Buller et al. 2011; Vaquero 2008).

Addressing the underlying challenges requires action that spans several tiers of archaeological knowledge production. Even if small, operational differences among individual researchers are persistent (Pargeter et al. 2023) and must be considered when proposing terminological unity. For this reason, it would be too optimistic to ask for terminological consistency at an individual level, given that definitions exist (Table 1). In our opinion, there is a greater need for transparency in reporting the excavation and post-excavation methods of defining SUs (e.g., Diez-Martín et al. 2023). Our proposal is to make the operational definition of contextual and stratigraphic units a peer-reviewed research output in its own right (e.g., Discamps et al. 2023). By doing so, the archaeological community can move toward a more coordinated framework in which researchers, reviewers, and editors collectively evaluate and refine replicable operational standards.

Finally, the need for standardization extends beyond individual excavations. Large-scale databases such as ROCEEH (https://hadw-bw.de/en/research/research-center/roceeh/digital-resources) and PaleoCore (paleocore.org) are building integrative platforms for archaeological data. However, their potential for cross-site comparison and cumulative knowledge building depends on consistent definitions of stratigraphic units. As the core of archaeological interpretation, contexts must be integrated into these infrastructures through interoperable vocabularies. Existing semantic infrastructures provide a natural foundation for this effort: formal ontologies such as the CIDOC Conceptual Reference Model, multilingual archaeological thesauri such as PACTOLS, and Linked Open Data initiatives already offer community-maintained frameworks into which explicitly defined stratigraphic vocabularies could be incorporated. Without such standardization, the promise of cumulative, comparative research in Paleolithic archaeology will remain unfulfilled. Our study suggests that NLP can play a role in this process, providing a replicable and transparent first step toward empirically grounded terminological frameworks that align methodological innovation with theoretical clarity.

Conclusion

The issues raised in this study underscore that stratigraphic units in Paleolithic archaeology are often treated as descriptive entities, their names functioning primarily as nominal markers that distinguish one unit from another in sequence. However, the rise of computational and spatial recording methods now enables us to conceive of archaeological sequences as multidimensional datasets—point clouds that can be interrogated through multiple variables. This shift challenges the traditional view of context as a static framework. Instead, the boundaries between geology, taphonomy, and human behavior increasingly blur, with each becoming one among many proxies used to define context.

Our results show that even when definitions exist, they are rarely applied consistently. Terms such as "layer," "level," "occupation," and "horizon" are frequently used interchangeably, with direct implications for interpretive outcomes. Inconsistent terminology is not a trivial matter: it risks producing different narratives from the same dataset. Although NLP provides a powerful means of detecting such inconsistencies, its role is diagnostic rather than interpretive. Models such as SciBERT allow us to map broad patterns of usage, but their insights must always be contextualized through close reading and critical engagement with archaeological questions.

Limitations also remain. SciBERT is trained on existing scientific literature and may reproduce disciplinary biases embedded in that corpus. For archaeology, a field characterized by regional traditions and varied theoretical frameworks, this is a significant constraint. Developing domain-specific models trained on archaeological texts could address this in the future, though such an effort would need to account for linguistic diversity, disciplinary background, and the prevalence of non-native English writing. The resources required for training may also outweigh the marginal gains in accuracy.

Moving forward, we suggest two paths. First, at the practical level, archaeologists should strive for consistency within individual publications by avoiding the use of multiple, overlapping terms for the same stratigraphic unit. Second, at the theoretical level, the field must confront the unexamined assumptions that have underpinned stratigraphy-related terminology since its inception. The problem is not only linguistic but epistemological. Terminology reflects theoretical commitments, and outdated or unchallenged frameworks continue to shape how we construct and interpret context. Addressing

these issues is essential if Paleolithic archaeology is to build cumulative, comparable knowledge while embracing the opportunities of new computational methods.

Acknowledgments. We thank the editor and the anonymous reviewers for their thorough and constructive comments across both rounds of review, which substantially improved the clarity and theoretical framing of this manuscript. AI-based language tools (ChatGPT, OpenAI) were used to assist with minor language editing and drafting of responses during manuscript preparation. All content was reviewed, verified, and approved by the authors, who take full responsibility for the final version of the manuscript.

Funding Statement. This research was funded by the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through a doctoral grant (UI/BD/150825/2021) and by the European Union (ERC-CoG-2021, FINISTERRA, 101045506). Open Access to this article is financed by Fundação para a Ciência e a Tecnologia (FCT), within the scope of the project ICArEHB UID/04211/2025 (https://doi.org/10.54499/UID/04211/2025).

Data Availability Statement. The dataset used in this study, consisting of published articles, is available on OSF: https://osf.io/sqcrn. The code used for data collection, processing, and analysis is available on GitHub: https://github.com/bili-piton/BERT-pipeline.

Competing Interests. The authors declare no competing interests.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/aap.2026.10181.
Supplementary Figure 1. UMAP of different contextual window sizes (*n_neighbors* = 50, windows = 10, 30, 40, 50).
Supplementary Figure 2. K-NN classification accuracy and macro-F1 across embedding window sizes (10–50).

References Cited

American Psychological Association. 2001. *Thesaurus of Psychological Index Terms*. 9th ed. American Psychological Association, Washington, DC.

Arteaga-Brieba, Andion, Lloyd A. Courtenay, Lucía Cobo-Sánchez, Antonio Rodríguez-Hidalgo, Palmira Saladié, Andreu Ollé, and Marina Mosquera. 2023. An Archaeostratigraphic Consideration of the Gran Dolina TD10.2 Cultural Sequence from a Quantitative Approach. *Quaternary Science Reviews* 309:108033. https://doi.org/10.1016/j.quascirev.2023.108033.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* 25(1):25–29. https://doi.org/10.1038/75556.

Bailey, Geoff. 2007. Time Perspectives, Palimpsests and the Archaeology of Time. *Journal of Anthropological Archaeology* 26(2):198–223. https://doi.org/10.1016/j.jaa.2006.08.002.

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv*. https://doi.org/10.48550/arXiv.1903.10676.

Bicho, Nuno, and João Cascalheira. 2020. The Use of Lithic Assemblages for the Definition of Short-Term Occupations in Hunter-Gatherer Prehistory. In *Short-Term Occupations in Paleolithic Archaeology: Definition and Interpretation*, edited by João Cascalheira and Andrea Picin, pp. 19–38. Springer International, Cham, Switzerland.

Binford, Lewis R. 1981. Behavioral Archaeology and the "Pompeii Premise." *Journal of Anthropological Research* 37(3):195–208. https://doi.org/10.1086/jar.37.3.3629723.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, California.

Blinkhorn, James, Clément Zanolli, Tim Compton, Huw S. Groucutt, Eleanor M. L. Scerri, Lucile Crété, Chris Stringer, Michael D. Petraglia, and Simon Blockley. 2021. Nubian Levallois Technology Associated with Southernmost Neanderthals. *Scientific Reports* 11(1):2869. https://doi.org/10.1038/s41598-021-82257-6.

Boas, Franz. 1932. The Aims of Anthropological Research. *Science* 76(1983):605–613. https://doi.org/10.1126/science.76.1983.605.

Boucher de Perthes, Jacques 1847. *Antiquités celtiques et antédiluviennes: Mémoire sur l'industrie primitive et les arts à leur origine*. Vol. 1. Treuttel et Wurtz, Paris.

Brandsen, Alex. 2023. Information Extraction and Machine Learning for Archaeological Texts. In *Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches*, edited by Cesar Gonzalez-Perez, Patricia Martin-Rodilla, and Martín Pereira-Fariña, pp. 229–261. Springer International, Cham, Switzerland.

Brandsen, Alex. 2024. *Archaeology Specific BERT Models for English, German, and Dutch. Zenodo*. https://doi.org/10.5281/zenodo.10610882, accessed June 3, 2025.

Brandsen, Alex, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2022. Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain. *Journal on Computing and Cultural Heritage* 15(3):51. https://doi.org/10.1145/3497842.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language Models Are Few-Shot Learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165.

Canals, Antoni, Josep Vallverdú, and Eudald Carbonell. 2003. New Archaeo-stratigraphic Data for the TD6 Level in Relation to *Homo Antecessor* (Lower Pleistocene) at the Site of Atapuerca, North-Central Spain. *Geoarchaeology* 18(5):481–504. https://doi.org/10.1002/gea.10071.

Clark, Amy E. 2017. From Activity Areas to Occupational Histories: New Methods to Document the Formation of Spatial Structure in Hunter-Gatherer Sites. *Journal of Archaeological Method and Theory* 24(4):1300–1325. https://doi.org/10.1007/s10816-017-9313-7.

Cunningham, Padraig, and Sarah Jane Delany. 2022. K-Nearest Neighbour Classifiers: A Tutorial. *ACM Computing Surveys* 54(6):1–25. https://doi.org/10.1145/3459665.

Dalmia, Ayush, and Suzanna Sia. 2021. Clustering with UMAP: Why and How Connectivity Matters. *arXiv*. https://doi.org/10.48550/arXiv.2108.05525.

Danilov, Gleb, Timur Ishankulov, Konstantin Kotik, Yuriy Orlov, Mikhail Shifrin, and Alexander Potapov. 2021. The Classification of Short Scientific Texts Using Pretrained BERT Model. In *Studies in Health Technology and Informatics*, Vol. 281, edited by John Mantas, Lăcrămioara Stoicu-Tivadar, Catherine Chronaki, Arie Hasman, Patrick Weber, Parisis Gallos, Mihaela Crişan-Vida, Emmanouil Zoulias, and Oana Sorina Chirila, pp. 607–611. IOS Press, Amsterdam.

Darvill, Timothy. 2003. *The Concise Oxford Dictionary of Archaeology*. Oxford University Press, Oxford.

Deschamps, Marianne, and João Zilhão. 2018. Assessing Site Formation and Assemblage Integrity Through Stone Tool Refitting at Gruta da Oliveira (Almonda Karst System, Torres Novas, Portugal): A Middle Paleolithic Case Study. *PLoS ONE* 13(2):e0192423. https://doi.org/10.1371/journal.pone.0192423.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805.

Diez-Martín, Fernando, Lucía Cobo-Sánchez, Cristina Fraile-Márquez, Sara De Francisco, Javier Duque-Martínez, Policarpo Sánchez-Yustos, Enrique Baquedano, Audax Mabulla, and Manuel Domínguez-Rodrigo. 2023. Searching for Intra-Site Spatial Patterns in the African Early Acheulean: The Lowermost Archaeo-Units at FLK West (Olduvai Gorge, Tanzania). *Archaeological and Anthropological Sciences* 15(5):64. https://doi.org/10.1007/s12520-023-01773-w.

Discamps, Emmanuel, Marc Thomas, Christelle Dancette, Brad Gravina, Sébastien Plutniak, Aurélien Royer, Alexandre Angelin, et al. 2023. Breaking Free from Field Layers: The Interest of Post-Excavation Stratigraphies (PES) for Producing Reliable Archaeological Interpretations and Increasing Chronological Resolution. *Journal of Paleolithic Archaeology* 6(1):29. https://doi.org/10.1007/s41982-023-00155-x.

Falcucci, Armando, Domenico Giusti, Filippo Zangrossi, Matteo De Lorenzi, Letizia Ceregatti, and Marco Peresani. 2024. Refitting the Context: A Reconsideration of Cultural Change Among Early Homo Sapiens at Fumane Cave Through Blade Break Connections, Spatial Taphonomy, and Lithic Technology. *Journal of Paleolithic Archaeology* 8(1):2. https://doi.org/10.1007/s41982-024-00203-0.

Farrar, Scott, and D. Terrence Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *Glot International* 7(3):97–100.

Fedele, Francesco G. 1976. Sediments as Palaeo-Land Segments: The Excavation Side of Study. In *Geoarchaeology: Earth Science and the Past*, edited by Donald A. Davidson and Myra L. Shackley, pp. 23–48. Duckworth, London.

Fedele, Francesco G. 1984. Towards an Analytical Stratigraphy: Stratigraphic Reasoning and Excavation. *Stratigraphica Archaeologica* 1:7–15.

Franzosi, Roberto, Wenqin Dong, and Yilin Dong. 2022. Qualitative and Quantitative Research in the Humanities and Social Sciences: How Natural Language Processing (NLP) Can Help. *Quality & Quantity* 56(4):2751–2781. https://doi.org/10.1007/s11135-021-01235-2.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *PNAS* 115(16):E3635–E3644. https://doi.org/10.1073/pnas.1720347115.

Garrod, Dorothy Anne Elizabeth. 1937. *The Stone Age of Mount Carmel Excavations at the Wady El-Mughara*. Vol. 1. Clarendon Press, Oxford.

Garrod, Dorothy Anne Elizabeth. 1942. *Excavations at the Cave of Shukbah, Palestine, 1928*. John Bellows Limited, Gloucester, UK.

Gasche, Hermann, and Onhan Tunca 1983. *Guide to Archaeostratigraphic Classification and Terminology: Definitions and Principles*. Ege Yayınları, Istanbul.

Gilbert, Allan S., and Paul Goldberg (editors). 2017. *Encyclopedia of Geoarchaeology*. Springer Reference, Dordrecht, Netherlands.

Goldberg, Paul, and Richard I. Macphail 2005. *Practical and Theoretical Geoarchaeology*. Blackwell Publishing, Oxford.

Hallinan, Emily, Omry Barzilai, Nuno Bicho, João Cascalheira, Yuri Demidenko, Mae Goder-Goldberger, Erella Hovers, Anthony Marks, Maya Oron, and Jeffrey Rose. 2022. No Direct Evidence for the Presence of Nubian Levallois Technology and Its Association with Neanderthals at Shukbah Cave. *Scientific Reports* 12(1):1204. https://doi.org/10.1038/s41598-022-05072-7.

Harris, Edward C. 1979. *Principles of Archaeological Stratigraphy*. Academic Press, London.

Hickman, Louis, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. 2022. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods* 25(1):114–146. https://doi.org/10.1177/1094428120971683.

Kipfer, Barbara Ann. 2021. *Encyclopedic Dictionary of Archaeology*. Springer International, Cham, Switzerland.

Kohavi, Ron. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2*, edited by Chris S. Mellish, pp. 1137–1143. Morgan Kaufmann, San Francisco.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings. *American Sociological Review* 84(5):905–949. https://doi.org/10.1177/0003122419877135.

Lartet, Édouard 1861. Nouvelles recherches sur la coexistence de l'homme et des grands mammifères fossiles réputés caractéristiques de la dernière époque géologique. *Annales des Sciences Naturelles, Zoologie* 15(3):177–253.

Lucas, Gavin. 2012. *Understanding the Archaeological Record*. Cambridge University Press, Cambridge.

Lyell, Charles. 1830–1833. *Principles of Geology*. John Murray, London.

Lyman, R. Lee, and Michael J. O'Brien. 1999. Americanist Stratigraphic Excavation and the Measurement of Culture Change. *Journal of Archaeological Method and Theory* 6(1):55–108. https://doi.org/10.1023/A:1022976726447.

Macanovic, Ana. 2022. Text Mining for Social Science: The State and the Future of Computational Text Analysis in Sociology. *Social Science Research* 108. https://doi.org/10.1016/j.ssresearch.2022.102784.

Machado Jorge, Hernández Cristo M, Mallol Carolina and Galván Bertila. 2013. Lithic Production, Site Formation and Middle Palaeolithic Palimpsest Analysis: In search of Human Occupation Episodes at Abric del Pastor Stratigraphic Unit IV (Alicante, Spain). *Journal of Archaeological Science* 40(5):2254–2273. https://doi.org/10.1016/j.jas.2013.01.002

Malinsky-Buller, Ariel, Erella Hovers, and Ofer Marder. 2011. Making Time: "Living Floors," "Palimpsests" and Site Formation Processes—A Perspective from the Open-Air Lower Paleolithic Site of Revadim Quarry, Israel. *Journal of Anthropological Archaeology* 30(2):89–101. https://doi.org/10.1016/j.jaa.2010.11.002.

Martín-Rodilla, Patricia. 2023. Computer Processing of Language: Where Archaeological Discourse and Computers Meet. In *Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches*, edited by Cesar Gonzalez-Perez, Patricia Martin-Rodilla, and Martín Pereira-Fariña, pp. 193–213. Springer, Cham, Switzerland.

Marwick, Ben. 2025. Is Archaeology a Science? Insights and Imperatives from 10,000 Articles and a Year of Reproducibility Reviews. *Journal of Archaeological Science* 180:106281. https://doi.org/10.1016/j.jas.2025.106281.

McInnes, Leland, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. https://doi.org/10.48550/arXiv.1802.03426.

McPherron, Shannon J. P., Harold L. Dibble, and Paul Goldberg. 2005. Z. *Geoarchaeology* 20(3):243–262. https://doi.org/10.1002/gea.20048.

Mielgo, Clara, Rosa Huguet, César Laplana, David M. Martín-Perea, Abel Moclán, Belén Márquez, Juan Luis Arsuaga, Alfredo Pérez-González, and Enrique Baquedano. 2024. Intra-Site Spatial Approaches Based on Taphonomic Analyses to Characterize Assemblage Formation at Pleistocene Sites: A Case Study from Buena Pinta Cave (Pinilla Del Valle, Madrid, Spain). *Archaeological and Anthropological Sciences* 16(1):5. https://doi.org/10.1007/s12520-023-01913-2.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Ning, Hui, and Zhenyu Chen. 2023. Fusion of the Word2vec Word Embedding Model and Cluster Analysis for the Communication of Music Intangible Cultural Heritage. *Scientific Reports* 13(1). https://doi.org/10.1038/s41598-023-49619-8.

Ojala, Markus, and Gemma C. Garriga. 2010. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research* 11(3):1833–1863.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, et al. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372:71. https://doi.org/10.1136/bmj.n71.

Pargeter, Justin, Alison Brooks, Katja Douze, Metin Eren, Huw S. Groucutt, Jessica McNeil, Alex Mackay, et al. 2023. Replicability in Lithic Analysis. *American Antiquity* 88(2):163–186. https://doi.org/10.1017/aaq.2023.4.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Perreault, Charles. 2019. *The Quality of the Archaeological Record*. University of Chicago Press, Chicago.

Pitt-Rivers, A. H. L. F. 1892. *Excavations in Cranborne Chase, Near Rushmore, on the Borders of Dorset and Wilts: Excavations in Bokerly and Wansdyke, Dorset and Wilts . . . with Observations on the Human Remains by J. G. Garson, M.D. 1892. Excavations in Cranborne Chase, Near Rushmore, on the Borders of Dorset and Wilts No. [Harrison and Sons, Printers].* Printed privately, London.

Renfrew, Colin, and Paul G. Bahn. 2004. *Archaeology: Theories, Methods and Practice*. 4th ed. Thames & Hudson, London.

Rousseeuw, Peter J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Saha, Rohan. 2023. Influence of Various Text Embeddings on Clustering Performance in NLP. *arXiv*. https://doi.org/10.48550/arXiv.2305.03144.

Schiffer, Michael B. 1972. Archaeological Context and Systemic Context. *American Antiquity* 37(2):156–165. https://doi.org/10.2307/278203.

Schiffer, Michael B. 1987. *Formation Processes of the Archaeological Record*. University of New Mexico Press, Albuquerque.

Shahack-Gross, Ruth. 2017. Archaeological Formation Theory and Geoarchaeology: State-of-the-Art in 2016. *Journal of Archaeological Science* 79:36–43. https://doi.org/10.1016/j.jas.2017.01.004.

Shott, Michael J. 1998. Status and Role of Formation Theory in Contemporary Archaeological Practice. *Journal of Archaeological Research* 6(4):299–329. https://doi.org/10.1007/BF02446082.

Sossa-Ríos, Santiago, Alejandro Mayor, Laura Sánchez-Romero, Carolina Mallol, Manuel Vaquero, and Cristo M. Hernández. 2024. The Time of the Stones: A Call for Palimpsest Dissection to Explore Lithic Record Formation Processes. *Journal of Archaeological Method and Theory* 31(4):2188–2238. https://doi.org/10.1007/s10816-024-09666-5.

Stein, Julie K., and Vance T. Holliday. 2020. Archaeological Stratigraphy. In *Encyclopedia of Geoarchaeology*, edited by Allan S. Gilbert, Paul Goldberg, Rolfe D. Mandel, and Vera Aldeias, pp. 1–7. Springer International, Cham, Switzerland.

Subakti, Alvin, Hendri Murfi, and Nora Hariadi. 2022. The Performance of BERT as Data Representation of Text Clustering. *Journal of Big Data* 9(1):15. https://doi.org/10.1186/s40537-022-00564-9.

Trigger, Bruce G. 2006. *A History of Archaeological Thought*. 2nd ed. Cambridge University Press, Cambridge.

Van der Maaten, Laurens, and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 9(86):2579–2605.

Vaquero, Manuel. 2008. The History of Stones: Behavioural Inferences and Temporal Resolution of an Archaeological Assemblage from the Middle Palaeolithic. *Journal of Archaeological Science* 35(12):3178–3185. https://doi.org/10.1016/j.jas.2008.07.006.

Vaquero, Manuel, María Gema Chacón, María Dolores García-Antón, Bruno Gómez de Soler, Kenneth Martínez, and Felipe Cuartero. 2012. Time and Space in the Formation of Lithic Assemblages: The Example of Abric Romaní Level J. *Quaternary International* 247:162–181. https://doi.org/10.1016/j.quaint.2010.12.015.

Cite this article: Galfi, Jovan, and João Cascalheira. 2026. Clarifying Stratigraphic Terminology in Paleolithic Archaeology Using Natural Language Processing. *Advances in Archaeological Practice.* https://doi.org/10.1017/aap.2026.10181.