WEP Warwick Evans Publishing

Transactions on Social Science, Education and Humanities Research
ISSN: 2960-1770, eISSN: 2960-2262 | Volume 16
EPHHR 2026

Lexical Stratigraphy and Turkicization in the Xiongnu: An Iranian Prestige Layer in Chinese Transcriptions

Shubin Gong *

Department of Linguistics, Emory University, Atlanta GA, United States

* Corresponding Author Email: sgong0528@gmail.com

Abstract. The Chinese historiographic record preserves ethnonyms, titles, personal names, and a modest set of glosses that later scholarship has treated as the principal basis for assigning a "Xiongnu language." This paper reexamines this long-standing Xiongnu language problem by treating the Chinese-recorded information as a stratified lexical record rather than as evidence for a single stable vernacular. Instead of assigning the material to one language family, the analysis focuses on the internal distribution of proposed etymological layers and their sociolinguistic implications. A clear asymmetry emerges: Iranian-associated items cluster in early sources and are largely confined to institutional and pastoral vocabulary, whereas Turkic-associated items appear across both early and later layers and span a broader range of semantic domains. This pattern is most consistent with an early Iranian prestige or contact layer progressively assimilated within a Turkic-dominant communicative environment during the Xiongnu imperial period. The findings suggest that apparent support for competing single-family hypotheses reflects stratification within a small, transmitted corpus rather than mutually exclusive identifications.

Keywords: Xiongnu; language contact; linguistic stratum; language shift.

1. Introduction

The Xiongnu occupy a pivotal position in Inner Asian history: they are central to early Chinese dynasties because of their sustained frontier and military presence yet persist as an interpretive problem because the surviving linguistic evidence is both small and methodologically compromised [1]. The Chinese historiographic record preserves ethnonyms, titles, personal names, and a modest set of glosses that later scholarship has treated as the principal basis for assigning a "Xiongnu language." Nevertheless, because the evidence consists almost entirely of Chinese transliterations, attempts to identify the language spoken by the Xiongnu have produced mutually incompatible proposals—Turkic, Iranian, Yeniseian/Paleo-Siberian, etc.—none of which has achieved broad consensus. Interpreting these transcriptions requires an explicit model of Old Chinese phonology; different reconstructions—such as the Baxter-Sagart reconstruction—yield different phonological correspondences and therefore different etymological affordances [2]. As Savelyev and Jeong note, the field's competing hypotheses are not merely a matter of interpretive preference; they arise because different subsets of the same small corpus appear to support different affiliations [3].

This paper accepts the limits of the available evidence—there is no extant continuous Xiongnu text comparable to the corpora that anchor historical phonology for better-attested families—and therefore shifts the analytic goal. Instead of attempting to assign the Xiongnu language to an attested family, the paper treats the lexical components in Chinese material as evidence for stratification and change within the confederation. The central question becomes historical and sociolinguistic: how could multiple etymological strata be embedded in a steppe confederation, and what does their distribution imply about language shift over time? The paper aims to answer this question by proposing that the Xiongnu lexicon as recorded in Chinese is best explained by an early Iranian prestige/contact stratum that becomes progressively assimilated within an increasingly Turkic-dominant environment during the imperial period as recorded in Han sources. "Turkic predominance," as used here, is strictly functional and distributional: it denotes the relative breadth and persistence of Turkic-associated

items across chronological layers and semantic domains in the recorded lexicon, not demographic homogeneity, ancestry proportions, or anachronistic projection from later Turkic empires.

2. Ethnolinguistic Identification as Evidence for Stratification

Xiongnu ethnolinguistic identification is often described as inconclusive. The persistence of competing hypotheses over more than a century is not simply the product of insufficient philological rigor; it is also an expected outcome when a small corpus preserves residues of multiple contact layers. Savelyev and Jeong review the competing proposals, noting that Turkic, Eastern Iranian, and Yeniseian hypotheses have all been advanced in modern scholarship [3]. The important observation is that these hypotheses typically do not explain the same pieces of evidence equally well; rather, they often track different subsets of a heterogeneous dataset.

If the available corpus reflects a single, stable language, one could expect the most transparent etymologies to cluster around one language family, with residual noise attributable to borrowing. Instead, the dataset repeatedly yields plausible proposals pointing in different directions. A stratified interpretation offers a near-satisfactory explanation for this otherwise puzzling fact: the Chinese material may preserve lexical and onomastic residues from multiple ethnolinguistic groups incorporated into the Xiongnu political space. Under such conditions, it is not surprising that different scholars, privileging different sources or weighting different reconstructions, reach different conclusions. The inferential task is, therefore, not to force a single-family solution onto a multi-source record, but to ask whether the distribution of strata shows historical directionality that can be related to ethnolinguistic dynamics within the Xiongnu confederation.

It does not propose new phonological reconstructions or a definitive identification of a single "Xiongnu language," nor does it attempt to adjudicate individual etymologies. Instead, it asks whether the proposed Iranian and Turkic layers exhibit distinct chronological and semantic profiles that would be expected under contact-induced stratification and language shift within a steppe confederation.

2.1. Archaeogenetic Constraints

Any claim about language shift within the Xiongnu must be plausible given the polity's demographic and social structure. In this regard, recent archaeogenetic studies do not "solve" the linguistic question, but they provide a crucial contextual constraint: they establish that heterogeneity was a long-run feature of Eastern Steppe populations and that the Xiongnu Empire was embedded in broader regional dynamics of mixture and mobility [4]. At the level of Xiongnu-period communities, Lee et al., analyzing genome-wide data from cemeteries on the western frontier, show that genetic diversity within local communities was comparable to that of the Empire as a whole, with substantial heterogeneity observed even within extended family groups [5]. They further argue that lower-status individuals exhibit the highest heterogeneity, whereas higher-status individuals are less diverse, implying that elite status was concentrated within specific genetic subsets of the broader Xiongnu population [5]. These findings substantiate a picture of the Xiongnu as a polity in which incorporation and mobility were routine and in which "Xiongnu" functioned as a political-collective label over genetically diverse constituencies.

Complementary work on earlier and contemporary Mongolian populations also supports repeated east–west mixture as a long-term feature of the region. Rogers and Kaestle, focusing on mitochondrial DNA haplogroup frequencies in the slab-burial mortuary culture (ca. 1100–300 BCE), report that western mitochondrial lineages were particularly prominent in Early Iron Age burial populations and note continuity in frequencies into Xiongnu-associated burial populations, which they take as supporting a genealogical relationship between slab-burial builders and Xiongnu-associated groups [6]. The broader implication is straightforward: western Eurasian genetic inputs were neither anomalous nor necessarily late intrusions. They form part of the demographic background in which an Iranian-speaking component could plausibly be present as a contact or incorporated stratum during the centuries in which the Xiongnu formed and expanded.

The relevance to language shift is sociological rather than directly probative of linguistic affiliation. A genetically heterogeneous polity with status-structured ancestry is the type of setting in which one expects multilingualism, elite-mediated prestige borrowing, and differential assimilation rates across groups. It is also the type of setting in which lexical strata may survive unevenly: prestige terms can be borrowed and retained long after the speech community that introduced them has shifted language, been absorbed, or disappeared. Genetics therefore supports the plausibility of a stratified linguistic record without being treated as evidence for any language family.

2.2. Lexical Stratigraphy in the Chinese-Recorded Xiongnu Corpus

The key evidence for the present argument is not any single etymology but the way proposed etymologies pattern across sources. In Savelyev and Jeong's interpretation of Dybo's analysis, Eastern Iranian and Turkic items show different chronological and semantic profiles [3, 7]. The Eastern Iranian set is concentrated in early (largely Western Han) material and is dominated by institutional vocabulary—especially political titles and terms connected with dairy pastoralism—with only one late-period common noun and one additional item glossed as 'comb' reported in Dybo's inventory [3]. By contrast, Turkic-associated items appear in both early and later layers and extend across a wider range of domains [3]. This distribution is more consistent with an early prestige/contact layer than with either random borrowing or an Iranian "core" vocabulary: titles and specialized pastoral terms are the kinds of items that can be borrowed from socially salient groups and persist as conservative residues after broader communicative practices shift.

Savelyev and Jeong further suggest that key components of Xiongnu political and economic organization may have been mediated by Eastern Iranian groups and that these Iranian-speaking communities were assimilated over time by a Turkic-speaking sector of the Xiongnu population [3]. The present paper adopts this directional reading but tightens two points that are often left underspecified in secondary discussion: the temporal locus of assimilation and the meaning of "predominant."

With respect to timing, the distribution itself already implies directionality within the period covered by Chinese sources. Eastern Iranian items cluster in earlier sources (especially Western Han), while Turkic items span both early and later attestations. This alone does not allow dating a discrete "shift event," but it does permit a cautious inference: by the time the earliest Western Han records preserve Xiongnu-linked terms, an Iranian-associated prestige/contact layer is visible, yet Turkic-associated material is already present and—crucially—persists into later layers. The reduction of Iranian-associated items in later sources is therefore most coherently read as a process of assimilation progressing during the imperial centuries rather than only after the polity's disintegration, marked by Western Han campaigns under Emperor Wu and subsequent political division within the confederation [1].

With respect to "predominant," it is essential to separate a functional inference from a demographic assertion. The data most securely supports predominance in the sense of functional and institutional range. Turkic-associated items appear across both early and late chronological layers and in a wider range of semantic domains, which is compatible with Turkic serving as a primary medium for intergroup communication and for maintaining a shared political vocabulary. This is precisely the sociolinguistic condition under which incorporated groups, including Iranian-speaking communities, would be expected to undergo language shift over time, while leaving behind institutionally entrenched lexical residues.

Therefore, the Iranian layer should be recognized as historically meaningful, but its distribution suggests that it is not the language of the entire polity. Titles and dairy-pastoral vocabulary are exactly the domains where an incorporated but prestigious group — specialists in pastoral production or actors embedded in elite coalitions—can exert linguistic influence disproportionate to their demographic weight. Meanwhile, the broader spread and persistence of Turkic-associated items suggests that Turkic was not simply a late addition but a durable stratum with long-term

communicative salience. The most plausible synthesis is therefore a Turkic-dominant environment that progressively absorbs Iranian-speaking groups during the imperial period, leaving early recorded Iranian terms as a prestige/contact residue.

This interpretation aligns well with the archaeogenetic constraints discussed above. Lee et al.'s demonstration of extreme heterogeneity at community and family scales, coupled with the concentration of elite status within specific subsets, provides a plausible social mechanism for the observed lexical pattern [5]. In a setting where political authority is exerted over heterogeneous groups, institutional vocabulary can be conservative, while everyday intergroup communication gravitates toward the language associated with the dominant political-military coalition. Under such conditions, prestige terms — especially titles — can retain older strata longer than other lexical domains, even as language shift proceeds.

2.3. Post-imperial Absorption and Successor Polities

Although the central claim concerns assimilation during the imperial period, post-imperial trajectories matter because they bear on plausibility: if Turkicization progressed within Xiongnu political space, communities emerging from that space could be re-embedded in successor formations without preserving clear Iranian linguistic dominance, while still retaining older lexical residues in narrow domains. North China's early medieval record illustrates how rapidly labels and lineages could be reconfigured under new regimes, providing a historical analogue for the kinds of sociolinguistic transitions implied by the Xiongnu lexical record.

Han and Later Han historiography presents the Xiongnu elite as an institutional system grounded in exogamous alliance and regional segmentation rather than as a single, internally uniform descent line. In the Hou Hanshu "Nan Xiongnu Liezhuan," the Chanyu is associated with the Luandi/Xulianti clan, while several clans—including Huyan, Xubu, Qiulin, and Lan—are identified as eminent clans that intermarried with the ruling line and occupied hereditary positions within the confederation [8]. This configuration not only attests to elite cohesion but also implies the incorporation of multiple high-status lineages whose origins are treated in the sources as distinct rather than subsumed under a single "Xiongnu proper" genealogy. Elite intermarriage and access to office thus functioned as mechanisms for integrating politically consequential groups of heterogeneous background into the ruling class, rendering elite-level linguistic plurality structurally plausible.

This institutional diversity is reinforced by evidence of post-imperial polities. Traditional commentaries note that some marriage-linked Xiongnu clans remained legible within later classificatory frameworks—for example, annotators gloss Huyan as corresponding to a contemporary Xianbei surname and observe the continued use of Lan as a lineage name [9]. Northern Dynasties historiography likewise records Xiongnu-linked elite houses within Xianbei-dominated political orders. These notices do not establish vernacular continuity, but they document that the elite houses associated with Xiongnu political space could be re-embedded across successive steppe and non-steppe regimes. Thus, persistence of clan names is best understood as evidence for elite transmission rather than proof of linguistic stability. It also aligns with Chinese scholarship documenting that Xiongnu-associated descent groups remained active after the empire's dissolution and subsequently entered the political configurations of later Turkic formations [10].

3. Discussion

The Xiongnu ethnolinguistic problem has often been treated as a classification task: to identify the language behind Chinese transliterations and to reconstruct a definitive "Xiongnu language." The Chinese-recorded corpus is small, chronologically layered, and likely to preserve residues from multiple groups incorporated into a steppe empire. Under these conditions, a stratified approach is not an evasion but a methodological necessity.

The decisive empirical observation remains the distributional asymmetry emphasized by Savelyev and Jeong: Iranian-associated items cluster in earlier sources and are semantically restricted, whereas

Turkic-associated items persist across early and late attestations and span wider semantic domains [3]. This asymmetry is difficult to reconcile with both an "Iranian core" model and a "random borrowing" explanation. It is, however, consistent with a process of Turkicization: early Iranian influence is visible in institutionally salient vocabulary, while Turkic functions as the more persistent and semantically expansive stratum, plausibly reflecting its long-term role in intergroup communication and political life. Read alongside archaeogenetic findings of extreme heterogeneity and status structure and evidence of substantial east–west genetic inputs in Xiongnu-period contexts, the sociolinguistic plausibility of such a shift is strengthened without conflating genes and language [5, 6].

The paper has intentionally limited its engagement with competing single-family hypotheses, beyond recognizing that their plausibility on subsets of the data is precisely what one expects under a layered record. Any model that seeks explanatory adequacy must account not only for a handful of attested etymologies but for the systematic contrast between early, semantically narrow Iranian-associated material and broader, temporally persistent Turkic-associated material. The stratified Turkicization account does so with fewer auxiliary assumptions than monolithic alternatives.

Several constraints and directions follow. First, the inference is only as reliable as the etymological classifications on which it rests. The argument therefore depends on the distributional robustness of classifications used by Savelyev and Jeong and earlier by Dybo [3,7]. It should be revisited as reconstructions of Old Chinese and Proto-Turkic are refined [2]. Second, the corpus would benefit from stricter source stratification: simple categories such as "early Xiongnu" versus "late Xiongnu" in Chinese historiography are too blunt, and future work should refine transmission layers.

4. Conclusion

Within the limits of present evidence, the most defensible conclusion is that the Xiongnu linguistic record preserved in Chinese is best approached as stratified. Its internal asymmetries are consistent with an early Iranian prestige/contact layer undergoing progressive assimilation in a Turkic-dominant communicative environment during the imperial period as recorded in Han sources. This remains an inference about sociolinguistic directionality, not a direct identification of a single vernacular. The main vulnerability is that this inference inherits the etymological assignments and source stratification used in Dybo's inventory and in Savelyev and Jeong's synthesis; if future work materially revises those classifications or the reconstruction of Chinese transcriptions, the distributional signal should be reassessed.

The contribution of this approach is methodological as much as substantive: it explains why mutually incompatible single-family identifications each can appear plausible on subsets of the same small corpus, and it reorients evaluation toward distributional predictions about strata (chronology and semantics) rather than toward isolated etymologies. Future work can strengthen or revise the account by refining the reconstruction through Chinese transliteration, expanding the inventory of securely attributable items, and reassessing etymological assignments under updated reconstructions of Old Chinese and proto-Turkic or other proto-languages.

References

[1] Di Cosmo N. Ancient China and Its Enemies: The Rise of Nomadic Power in East Asian History. Cambridge: Cambridge University Press, 2002.

[2] Baxter W H, Sagart L. Old Chinese: A New Reconstruction. Oxford: Oxford University Press, 2014.

[3] Savelyev A, Jeong C. Early nomads of the Eastern Steppe and their tentative connections in the West. Evolutionary Human Sciences, 2020, 2: e20. DOI: 10.1017/ehs.2020.18.

[4] Jeong C, Wang K, Wilkin S, et al. A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. Cell, 2020, 183(4): 890-904.e29. DOI: 10.1016/j.cell.2020.10.015.

[5] Lee J, Miller B K, Bayarsaikhan J, et al. Genetic population structure of the Xiongnu Empire at imperial and local scales. Science Advances, 2023, 9(15): eadf3904. DOI: 10.1126/sciadv.adf3904.

[6] Rogers L L, Kaestle F A. Analysis of mitochondrial DNA haplogroup frequencies in the population of the slab burial mortuary culture of Mongolia (ca. 1100–300 BCE). American Journal of Biological Anthropology, 2022, 177(4): 644-657. DOI: 10.1002/ajpa.24478.

[7] Dybo A V. Lingvističeskije kontakty rannix tjurkov. Leksičeskij fond. Pratjurkskij period [Linguistic contacts of the early Turks. Lexical stock. The Proto-Turkic period]. Moscow: Vostochnaya literatura, 2007.

[8] Fan Ye. Hou Hanshu. Beijing: Zhonghua Shuju, 1965.

[9] Sima Qian. Shiji (Annotated with traditional commentaries). Beijing: Zhonghua Shuju, 1959.

[10] Wei L, Li H. About the names of the Chanyu family and the branch tribes of the Xiongnu. In: Xu D, Fu J (Eds.), Language Contact and Language Variation, pp. 69-96. Beijing: The Commercial Press, 2019.