From Altaic to Transeurasian: Traditional Genealogical Classification and Algorithmic Clustering

Authors: Jiang Di, Jiang Di, Wen Meng, Jiang Di, Wen Meng

Date: 2026-04-15T13:10:49+00:00

Abstract

By employing outgroup settings and rooting techniques in clustering experiments, this article performs technical judgments on the ingroup or outgroup status of the Ural-Altaic, traditional Altaic, and the recently proposed Transeurasian language (family) hypotheses, ultimately concluding that such related hypotheses are difficult to verify. The experiments not only conduct cross-calculations of similarity distances between languages and language groups for over a hundred languages or dialects across the Indo-European, Uralic, Altaic, Caucasian, Koreanic, Japonic, Sino-Tibetan, and Austronesian families, but also utilize Sinitic dialects, the Uto-Aztecan family of North America, and the Dravidian family of South Asia as outgroups to determine evolutionary sequences. Regarding the specific computational models, the Levenshtein distance method and related phylogenetic clustering software (MEGA) were applied. In terms of implementation, language quantity parameters and radical experimental schemes were proposed to facilitate logical judgments of the experimental results. Ultimately, the objective calculations refute the Ural-Altaic and Transeurasian language (family) hypotheses and fail to confirm the traditional Altaic language family. Concurrently, the study proposes the potential existence of a phylogenetically cognate language family in Northeast Eurasia: the Mongolic-Tungusic language family (Mongol-Manchu family).

Full Text

Preamble

From Altaic to Transeurasian: Traditional Genealogical Classification and Algorithmic Clustering

Jiang Di[1,2], Meng Wen[3]

(1. Laboratory of Language Sciences, Jiangsu Normal University, Xuzhou 221009; 2. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081;

3．人民教育出版社，北京 100081）

Abstract

This study employs outgroup settings and rooting techniques in clustering experiments to evaluate the validity of the Ural-Altai, traditional Altai, and the recently proposed Transeurasian language family hypotheses. By assessing the inclusion or exclusion of specific language groups, the research concludes that these hypotheses are difficult to substantiate. The experiments involve calculating similarity distances across hundreds of languages and dialects from the Indo-European, Uralic, Altaic, Caucasian, Koreanic, Japonic, Sino-Tibetan, and Austronesian families. To determine evolutionary sequences, Chinese dialects, the Uto-Aztecan family of South America, and the Dravidian family of South Asia were utilized as outgroup controls.

The computational model utilizes Levenshtein distance metrics processed through phylogenetic clustering software (MEGA). The experimental design incorporates language quantity parameters and radical experimental protocols to facilitate logical judgment of the results. Ultimately, the objective computational analysis refutes the Ural-Altai and Transeurasian hypotheses and fails to confirm the traditional Altai language family. Concurrently, the study proposes the potential existence of a genetically related language family in Northeast Eurasia: the Mongolic-Tungusic (Mongol-Manchu) language family.

关键词

Clustering Experiments, Levenshtein Distance, Quantitative Comparative Experiments, Altaic Languages, Transeurasian Languages

Transeurasian Languages and the Early Altaic Hypothesis

The vast Eurasian steppe, stretching from the Black Sea to the Far Eastern reaches of Siberia, has served for millennia as a cradle for countless ethnic groups that have emerged, vanished, or survived to the present day. This region has also been a crucible for the growth, convergence, development, and extinction of diverse languages. Among these, a legendary and persistently debated narrative exists: beginning with the ancient Ural-Altaic hypothesis, it was later narrowed down to the Altaic family, eventually expanding into the Macro-Altaic hypothesis—which includes Korean and Japanese—and most recently evolving into the Transeurasian language hypothesis (Wu 1996; Fei 2024:41).

Regarding the classification of Altaic languages, researchers have traditionally employed the classical historical-comparative method. However, numerous controversies remain, and the crux of the issue consistently lies in the difficulty of

determining whether similar linguistic features result from common ancestry or lexical borrowing (Poppe 2004/1965; Miller 1971). This distinction has proven difficult to clarify effectively. By examining various early classification perspectives on the regional languages, this paper adopts the Levenshtein distance model to calculate linguistic relationships and obtain data on inter-linguistic similarity. We propose a non-historical comparative approach to investigate this hypothesis, termed the quantitative comparative experimental method, and utilize clustering algorithms to verify the relevant propositions.

This work is a concluding result of the National Natural Science Foundation of China project "Origins and Evolution of East Asian Populations and Languages from a Global Linguistic Perspective" (31271337).

This research was also supported by the National Social Science Fund of China projects "Research on the Development and Construction of an Online Retrieval System for Large-scale Grammatically Annotated Texts of Chinese Ethnic Languages" (21&ZD304) and "Historical Comparative Linguistic Research of Sino-Tibetan Based on Large-scale Lexical and Phonetic Databases" (12&ZD174). This article is an expanded version of a paper with the same title published in *Contemporary Linguistics* (2026, Issue 2). The expanded content includes a comprehensive review of the concepts, backgrounds, and recent research progress regarding the Altaic and Transeurasian language families. We would like to express our gratitude to the anonymous reviewers and the editorial department of *Contemporary Linguistics* for their valuable revision suggestions, and to Dr. Yan Haixiong for his suggestions regarding the revision of the article title.

Provided valuable revision suggestions.

1.1 命题的渊源和复杂性

Transeurasian is a relatively new term introduced into academic discourse over the past two decades (Robbeets 2006, 2021) to describe the relationship between languages across the Trans-Eurasian region. Specifically, this grouping includes the Turkic, Mongolic, and Tungusic languages, as well as Korean and Japanese. From the perspective of academic history, the concept of Transeurasian is rooted in the classificatory framework of the Altaic language family.

The Altaic language family refers to a group of trans-regional languages—including Turkic, Mongolic, and Tungusic—that were observed to possess similarities as early as the 18th century. The term "Altaic" was first proposed by the Finnish linguist Matthias Castrén; however, his definition of the family also encompassed Finno-Ugric and Samoyedic, forming what would later be known as the "Ural-Altaic" hypothesis (Chisholm 1911). Several years later, the Austrian scholar Anton Boller (1811–1869) suggested the inclusion of Japanese into the Ural-Altaic family. In the 20th century, the Finnish linguist G. J. Ramstedt (2004/1957), often hailed as the "father of comparative Altaic linguistics," systematically organized these hypotheses. Ramstedt proposed a comprehensive Altaic hypothesis that rejected the Ural-Altaic connection while formally incor-

porating Korean into the Altaic family—a view that gained the support of most Altaicists (Ramstedt 1924). In 1960, Nicholas Poppe (1965), another major figure in Altaic studies, published a substantial revision of Ramstedt' s work on phonology, which has since become the standard reference for Altaic research. It can be said that Ramstedt and Poppe were the primary proponents of the modern Altaic theory. The theory was further expanded when R. A. Miller (1971) published *Japanese and the Other Altaic Languages*, which convinced many scholars that Japanese also belonged to the Altaic family.

Another term worth mentioning is the "macro-family" model, which was used when the early concept of "Ural-Altaic" was prevalent in academia. The Russian scholar Sergei Starostin (2003), one of the most steadfast proponents of the Altaic theory, also employed the concept of a "Macro-Altaic"family. However, his version of the macro-family included only Turkic, Mongolic, Tungusic, Korean, and Japanese, while excluding the Uralic languages. This scope represents the consensus generally accepted by later Altaic linguists.

Clearly, the connotations of these two terms have shifted throughout the history of Altaic classification. Beyond these views, there are entirely different perspectives on the classification of Altaic languages. For instance, Joseph Greenberg (2000) constructed an even broader macro-family concept called "Eurasiatic," which included Turkic, Mongolic, and Tungusic, as well as Korean, Japanese, and Ainu, alongside the traditional Indo-European and Uralic families (cf. Jiang 2000, 2012). Conversely, J. Marshall Unger (1990) proposed a "Macro-Tungusic" family consisting of Tungusic, Korean, and Japanese, while arguing that Turkic and Mongolic should be classified as independent language families.

These prior studies extensively cover the classification and boundaries of the Altaic family. Simultaneously, research regarding the "Urheimat" or linguistic homeland of the Altaic family has been frequently discussed. The predominant view remains the early nomadic culture hypothesis, which posits that the historical Altaic peoples gathered on the Central Asian steppes and originated in the Altai Mountains spanning Russia, Kazakhstan, and China. They are thought to have migrated westward and eastward to form the Western Turkic, Central Mongolic, and Eastern Tungusic peoples, respectively. Regarding the divergence time of Proto-Altaic, Ramstedt (2004/1957) suggested it occurred around 4000–3000 BCE; that is, by the early Bronze Age, Proto-Altaic had already split into the Turkic, Mongolic, and Tungusic branches. However, scholars such as Gerard Clauson (1959, 1960) argued that because much of the shared Altaic vocabulary consists of loanwords, the interference from these borrowings makes it impossible to trace the divergence back to the Neolithic period.

For convenience, abbreviated names are used for language families, groups, and individual languages where necessary in the text, particularly in discussions of topological structures. These include Mongolic (Mon.), Korean (Kor.), Japanese (Jap.), Tungusic (Tun.), Turkic (Turk.), Altaic (Alt.), Uralic (Ural.), Sino-Tibetan (ST.), and Austronesian (AN.), among others. Korean and South Korean (Hanguo-yu) are treated as the same language and are collectively re-

ferred to as Korean.

Webpage: BOLLER, Anton: https://whowaswho-indology.info/862/boller-johann-anton/. Accessed June 27, 2025.

Through lexical comparison and phonetic correspondences in the *Etymological Dictionary of the Altaic Languages*, Sergei Starostin (2003) determined the divergence time of Proto-Altaic to be approximately 5000–4000 BCE, a period close to the early stages of agricultural expansion.

1.2 新术语：泛欧亚语言

Transeurasian is a new term proposed in the 21st century, encompassing five major language families (or groups): Turkic, Mongolic, Tungusic, Koreanic, and Japonic. Among these, the Turkic, Mongolic, and Tungusic groups are referred to as Altaic in the narrow sense, while the inclusion of Koreanic and Japonic constitutes the Transeurasian (macro-family) concept. Clearly, this new concept aligns fundamentally with the "Macro-Altaic" framework proposed by the Russian scholar Sergei Starostin. Consequently, the new school represented by Martine Robbeets, who coined the term, does not exceed the boundaries of traditional views regarding linguistic classification. What, then, is the primary difference between the two?

Robbeets and her team do not limit themselves to traditional comparative methods. Instead, they actively employ statistical and computational models —specifically Bayesian phylogenetic algorithms—to perform automated analysis of large-scale datasets, while simultaneously emphasizing the study of contact, borrowing, and fusion between languages. This approach allows them to reconstruct the evolutionary history of these languages. They have also conducted in-depth structural analyses of the striking typological similarities across the Transeurasian languages, such as vowel harmony, agglutinative morphology, and Subject-Object-Verb (SOV) word order, providing detailed arguments for each feature. Regarding lexical and morphological similarities, while they face pressure to exclude loanwords, they do not entirely dismiss the possibility of these features originating from a common ancestral language. Finally, by integrating cross-disciplinary evidence from linguistics, archaeology, and genetics, they trace and reconstruct the historical migrations, fusions, and cultural connotations of the populations speaking these languages, using computational inference to determine their place of origin and dates of divergence. Their conclusion is that the prehistoric Transeurasian language family originated among millet farmers in North China. This refutes the earlier "nomadic origin" hypothesis associated with the Eurasian Steppe and brings a brand-new historical perspective, tracing the potential center of origin for all Transeurasian languages to early Neolithic agricultural populations in Northern China (Robbeets et al., 2021). We recognize that the concept of Transeurasian is essentially a re-verification and reshaping of the Altaic hypothesis. From its inception in the 18th century until the 1950s, the Altaic theory remained highly controversial as a classification

hypothesis and was eventually largely abandoned. The fundamental reason for this was the inability to verify cognates within the family and the failure to discover hypothesized laws of sound change. Conversely, evidence suggested that the three main groups—Turkic, Mongolic, and Tungusic—tended toward similarity due to long-term mutual contact and regional influence. This resulted in the formation of a "Sprachbund" (linguistic area) relationship, which simultaneously negated their genetic relationship.

The Transeurasian language family is one of the few in the world with a transnational and trans-continental distribution. From a purely geographical standpoint, it appears to be a simple distribution connecting the Eurasian landmass. However, from the perspective of human civilization, it represents a grand historical narrative of the flourishing of nomadic civilizations across the Eurasian Steppe and the dissemination of agricultural culture and technology over tens of thousands of years. Objectively speaking, Transeurasian research is the product of highly complex interdisciplinary work; whether the conclusions of the Robbeets team are reliable and credible is often difficult for non-specialist readers to judge. As for whether the Transeurasian languages were indeed spread and facilitated by the early millet farming civilizations of Eastern Eurasia (specifically the West Liao River basin in Northeast China), further research is required for verification.

Reviewing the history of related research, Transeurasian studies have undergone three key stages, which can be represented by three major works from Robbeets and her team. The first stage involved the proposal and establishment of the hypothesis, represented by her 2005 doctoral dissertation, *Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?*. The second stage deepened the research using the verbal morphology of the five major language groups, resulting in the 2015 technical monograph, *Diachrony of Verb Morphology: Japanese and the Transeurasian Languages*. The third stage began the construction of a theoretical framework, which utilized contact linguistics to distinguish between cognate morphology and regional diffusion, and introduced Bayesian phylogenetics to study computational divergence dates. In 2020, this school' s collective achievements were compiled into the *Oxford Guide to the Transeurasian Languages*.

Understanding this makes it clear why, even as the Altaic hypothesis was being gradually abandoned by many, Robbeets' "Transeurasian Roots" project was still able to secure funding from the European Research Council (ERC) (a 2.5 million Euro grant approved in 2019). It also highlights the historical academic value of the project and its profound potential to reveal the origins of human civilization and the foundations of contemporary global diversity.

This series of studies facilitated a paradigm shift in the team' s research. Beyond a purely linguistic perspective, they integrated archaeological and genetic considerations. This culminated in the major 2021 study published in *Nature*, "Triangulation supports agricultural spread of the Transeurasian languages." This work combined linguistic, archaeological, and genomic evidence to support the

central thesis that the expansion and diversification of Transeurasian languages were likely driven by the spread of agriculture (Robbeets et al., 2021).

The research by Robbeets et al. has garnered extensive international attention, and their published works have received numerous academic reviews. Two significant reviews are highlighted here. The first is by the renowned linguist Geoffrey Sampson (2022) regarding *The Oxford Guide to the Transeurasian Languages*. Sampson notes that the collection possesses considerable breadth and depth, serving as a comprehensive and detailed reference for Transeurasian languages, and that the historical and structural descriptions of the five language families are of high value. However, he argues that the evidence for the Transeurasian macro-family hypothesis remains flawed and too weak to convince skeptics. The primary issue remains the lack of rigorous demonstration of shared features, such as vowel harmony and SOV word order.

The second review is titled "Comment: Triangulation fails when linguistic, genetic and archaeological data do not support the Transeurasian narrative" (Zheng, Tian, et al., 2022). This critique was produced by a collaborative team of linguists, geneticists, and archaeologists. The reviewers argue that all three components of the original paper suffer from serious problems. In linguistics, for instance, while Robbeets et al. provided 3,166 cognate sets as evidence, the reviewers claim these data do not adhere to the core principles of phonetic correspondence in historical linguistics (including the exclusion of loanwords and chance similarities). The reviewers point out: "A systematic computer-aided analysis of the remaining cognate sets shows that only 17 etymologies meet the authors' own criteria for identifying regular sound correspondences. This means that out of 3,166 cognate sets, only 17 potentially support the hypothesis ···" They conclude that the linguistic evidence is insufficient to distinguish between chance similarity, contact, and inheritance, and thus does not support the claims of Robbeets et al. Furthermore, the genetic and archaeological sections also face severe issues. For example, the original paper claimed that West Liao River farmers (with mixed Yellow River farmer ancestry) spread agriculture to Korea and Japan, thereby driving language dispersal. However, re-analysis of the data suggests that the original genetic model was highly selective, ignored alternative hypotheses, and could not distinguish between different ancestral sources, rendering the original hypothesis untenable. Finally, the review concludes that the "Triangulation" paper provides neither conclusive evidence for a Transeurasian language family nor evidence linking five distinct language families to the Neolithic dispersal of farmers from the West Liao River region. The Altaic hypothesis is an old topic that has undergone extremely complex internal reconstructions and external expansions by generations of scholars, with the content of the hypothesis constantly changing. Robbeets has attempted to reshape the Altaic—or Macro-Altaic—family using new technological developments. Although the name has changed to "Transeurasian," the underlying objective remains the same.

Taking advantage of this opportunity in the contemporary era of the internet and

large-scale linguistic data, this paper intends to examine this ancient hypothesis through new technical methods, aiming to reach new conclusions and advance our understanding of the subject.

The End of the Ural-Altaic Hypothesis

2.1 乌拉尔-阿尔泰语系争讼之背景

Through the preceding review of the research history of the traditional Altaic language family, we have noted several significant concepts that have emerged, such as the so-called "Ural-Altaic language family hypothesis." We believe it is precisely these valuable ideas that have guided Altaic studies for over a century, with an influence that extends to the present day.

The senior corresponding authors of the team include Thomas Pellard (Linguistics), Chuan-Chao Wang (Genetics), and Shaoqing Wen (Archaeology). Additionally, the prominent scholar Guillaume Jacques, a specialist in Sino-Tibetan languages from France, is also a member.

Article Title: "Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative Arising from Robbeets et al." Published in *bioRxiv* preprint doi: https://doi.org/10.1101/2022.06.09.495471; this version posted June 12, 2022.

As this article was being finalized, Robbeets et al. published a new book, *The Oxford Guide to the Archaeology of Ancient Language Change* (2025); please refer to the postscript of this article for further details.

The Transeurasian linguistic theory is a significant framework; however, whether the hypothesis it reflects represents a common genetic origin or merely a relationship of contact requires scientific demonstration. This section intends to examine this hypothesis using a non-historical comparative method referred to as a quantitative comparative experiment. Before conducting the experiment, it is necessary to briefly understand the current state of Uralic language studies and their potential relationship with the Altaic languages. According to early linguistic historical records, the Uralic languages originated near the Ural Mountains on the East European Plain (in present-day Russia) approximately 7,000 to 4,000 years ago. This language family gradually differentiated due to the migration of original hunter-gatherer groups. One branch moved north along the upper reaches of the Volga River into Northern Europe to form the Finno-Ugric branch, producing the northern languages of this group, such as Finnish, Estonian, and Saami. The eastern branch crossed the Ural Mountains into Siberia, forming the Samoyedic languages, including modern Nenets and Selkup. A westward branch formed the Magyars, which further differentiated into Hungarian (which entered Central Europe) and languages scattered throughout the Volga basin and near the Ural Mountains, such as Komi, Mari, and Udmurt. The Uralic languages expanding westward and southwestward must have come into close contact with the Indo-European language family. Simultaneously, Tur-

kic languages continuously expanded from south to north throughout history, interacting with both Indo-European and Uralic languages, particularly in the Greater Caucasus region spanning Eurasia (the North Caucasus, primarily in Russia) and the South Caucasus (including Georgia, Armenia, and Azerbaijan). In this sense, it is entirely possible that Uralic languages developed a convergent relationship with neighboring languages through linguistic contact [**?**].

At the boundary between Europe and Asia and within the regions of the Caspian Sea, Black Sea, and Mediterranean, there is a degree of geographical overlap and proximity between the Uralic and Altaic language families, particularly with the Turkic branch. From a micro-perspective, languages of these two families do indeed share certain common features, such as universal agglutinative characteristics, rich case systems (15 cases in Finnish and 24 in Hungarian), vowel harmony, and some lexical similarities. Consequently, whether the relationship between the two is one of distant common ancestry or proximity-induced convergent similarity has been a subject of controversy for over a century. Roughly after the 1940s, N. Poppe [**?**] attempted to prove the genetic relationship of these languages by searching for regular phonetic correspondences, shared roots, and grammatical inheritance. Subsequently, research by scholars such as D. Sinor [**?**] and S. Starostin [**?**] concluded that there is no phylogenetic relationship between the Uralic and Altaic language families. As a result, the Ural-Altaic hypothesis has had almost no firm supporters since the 1960s.

2.2 实验语料预览

This study employs the Levenshtein distance algorithm to calculate the similarity relationships between the Uralic and Altaic language families. To assist in the analysis, neighboring Indo-European languages are included as a reference. Using phonetically transcribed Swadesh 100-word lists, this research attempts to determine whether the observed similarities between these families are the result of common genetic descent or linguistic contact.

The linguistic data used in the experiments are categorized by language family. To ensure the reliability of the experimental results, two sets of languages were prepared for each family to conduct reinforced parallel experiments. The languages of the Altaic, Japonic, and Koreanic families are organized as follows.

We previously considered a hypothesis linking Europe with the Ewenki people of China through shamanism and reindeer culture—specifically, whether the relationship between Uralic and Altaic languages might be related to the migration of circumpolar reindeer-herding peoples and the subsequent dispersal of their languages. Under this framework, a high-latitude circumpolar corridor—encompassing Sweden, Norway, Finland, Canada, the United States (Alaska), the Russian Far East (including various Sámi groups), and the Greater Khingan Range in China (Ewenki groups)—could have facilitated the spread or connection of these languages. At present, however, this idea appears to rely too heavily on intuition, and the burden of proof remains exceedingly high.

In terms of clustering techniques, Indo-European languages were originally selected as an outgroup. The term "auxiliary judgment" is a deliberate ambiguity; we treat Indo-European as part of the ingroup to observe its potential relationship with Uralic. Logically, treating Indo-European as an ingroup contradicts the traditional Ural-Altaic hypothesis. However, from another perspective—following Greenberg' s (2000) Eurasiatic theory—Indo-European can coexist with Uralic and Altaic within the same macro-system. In this sense, the operation in this section serves as a technical processing strategy.

The languages are listed in groups, though Indo-European and Uralic are not subdivided. Another purpose of this list is to help readers identify the positions of specific languages in the experimental diagrams; therefore, the following list includes the English or Hanyu Pinyin names of the languages in parentheses.

Language names in the experimental diagrams are constructed according to the following rules: the first abbreviation represents the language family, and the second represents the branch, sub-branch, or language group. These are followed by an underscore to separate the language from its higher-level classification. Language names are written in lowercase Pinyin or English. If a language includes multiple survey points or dialects, the first letter of the language name is capitalized after the underscore, followed by the dialect name in lowercase. The "1" at the end of the filename is a technical requirement for data processing. Only the necessary family and branch abbreviations are listed below: IG (Indo-European: Germanic), IR (Romance), IC (Celtic), IS (Slavic), IB (Baltic), UF (Uralic: Finnic), UU (Ugric), UV (Volga-Permic), US (Samic), UD (Samoyedic), UL (Mari), UM (Mordvinic), AT (Turkic), AU (Tungusic), AM (Mongolic), KO:

Koreanic; JP: Japonic; AN: Austronesian; CAU: Caucasian. Some languages were used only in preliminary experiments and are not listed here.

Turkic Family. Group A: Turkish, Turkmen, Kazan Tatar, Nogai, Kumyk. Group B: Bashkir, Kazakh, Azerbaijani, Crimean Tatar, Uyghur.

Mongolic Family. Group A: Bonan (Nianduhu dialect), Eastern Yugur, Daur (Meilisi dialect), Buryat, Mongolian (Zhenglan Banner dialect).

Group B: Mongolian (Darhan dialect), Daur (Nierji dialect), Bonan (Baoancheng dialect), Dongxiang (Ili dialect), Tu/Monguor (Minhe dialect). Group C: Kangjia, Tu/Monguor, Yugur (Kangle dialect), Dongxiang, Mongolian (Dulan dialect), Mongolian (Alasha dialect), Mongolian (Chen Barag dialect), Mongolian (East Sonid dialect), Mongolian (Harqin dialect), Mongolian (Bairin Right Banner dialect).

Tungusic Family. Group A: Orok, Orochi, Ulchi, Even, Negidal. Group B: Hezhen (Nanai), Nanai, Oroqen, Evenki, Manchu.

Japonic Family. Group A: Ainu (Hokkaido), Ryukyuan (Amami), Japanese (Kagoshima), Japanese (Hokkaido), Japanese (Kyoto).

Group B: Ainu (Sakhalin), Ryukyuan (Miyako), Japanese (Tokyo), Japanese

(Fukuoka), Japanese (Hachijo).

Koreanic Family: Pyongyang dialect, Seoul dialect, Jeollado dialect, Jeju dialect (Jeju language), Yanbian dialect (Jilin, China). Uralic Family: Hill Mari (Mari branch), Eastern Mari (Mari branch), Moksha (Mordvinic branch), Udmurt (Permic branch), Erzya (Komi branch), Hungarian (Ugric branch), Northern Mansi

Korean and Japanese are traditionally considered language isolates or primary language families. In practice, some scholars argue that the Jeju dialect in South Korea has evolved into the independent Jeju language. The Japonic family is currently recognized as including Japanese, Ryukyuan, and Ainu, among others.

The primary data sources for this paper are as follows: Portions were extracted from the Wiktionary Appendix of Swadesh lists (URL: https://en.wiktionary.org/wiki/Appendix:Swadesh_{list accessed October 8, 2024). Other portions were taken from Sun Hongkai, Ting Pang-hsin, Jiang Di, and Yan Haixiong (Eds.), *Phonology and Lexicon of Sino-Tibetan Languages*; the "Language Resources Protection Project of China" (URL: https://zhongguoyuyan.cn/index.html?lang=cn); Sun Zhu (Ed.), *Dictionary of Mongolic Languages*; Chen Zongzhen (Ed.), *Lexicon of Turkic Languages in China*; and the Tungusic section of Volume 6 of Sun Hongkai (Ed.), *Brief Surveys of Minority Languages in China* (Revised Edition). Additionally, Dr. Wang Haibo from Guangdong provided word lists for Orok and Nanai, for which we express our gratitude.

While Korean and Japanese are generally regarded as single languages with various dialects, a newer perspective suggests that both are languages distributed across geographically and maritimately isolated regions. See Ran Qibin and Weixiyu (2018).

(Northern) Mansi (Ugric branch), Karelian (Finnic branch), Estonian (Finnic branch), Finnish-Suomi (Finnic branch).

Indo-European Family. Germanic branch: English, German. Celtic branch: Welsh. Romance branch: Catalan, Spanish, Romanian. Slavic branch: Russian, Polish, Ukrainian, Serbian. Independent branches: Armenian, Greek, Albanian.

Caucasian Languages: Laz, Svan, Mingrelian, Georgian.

Certain sections utilize Austronesian and Sino-Tibetan languages for comparative analysis. Austronesian Family. Group A: Indonesian, Sundanese, Cebuano, Madurese, Kavalan. Group B: Malay, Amis, Tongan, Tagalog, Javanese. Sino-Tibetan Family: Tibetan (Hongyuan dialect), Yi (Xide dialect), Cantonese (Guangzhou), Xi' an dialect, Taiyuan dialect.

Outgroup languages, with Austronesian further divided into groups P and Q: Austronesian. Group P: Hawaiian, Fijian, Tahitian. Group Q: Amis, Tongan, Tagalog.

Sinitic: Jinan dialect, Taiyuan dialect, Xi' an dialect. The latter two are also used as representatives of the Sino-Tibetan family.

Americas, Uto-Aztecan Family: Nahuatl, Tarahumara.

South Asia, Dravidian Family: Tamil, Telugu, Malayalam.

2.3 借用特征在聚类中的独特作用

In this study, several preliminary experiments were conducted regarding the selection of the number of languages. These resulted in two primary clustering outcomes based on language count: Type A, which aggregates more than 20 languages, and Type B, which consists of fewer than 20 languages. It should be noted that the language count here is merely an estimate; the underlying cause lies in the borrowing features and the volume of borrowing cases generated by the proximity of related languages. Some sets with fewer than 20 languages may still produce Type A clusters if they contain a high density of complex borrowing features. Given that this paper primarily utilizes computational models for discussion, it does not intend to dwell on the minutiae of specific linguistic features. The following macro-analysis and narrative are framed around the topological and evolutionary structures (rooting) presented by the experimental results. Topology, in this context, refers to an analytical method for clustering relationships in linguistic data. Its significance lies in providing a perspective that transcends traditional statistical methods, focusing on connections, cycles, or breakpoints (holes) between languages, and particularly on the global structure of the data.

It captures the inherent geometric properties of linguistic data—much like observing the layout of mountains and rivers on a map—thereby revealing hidden linguistic relationships and patterns. For example, consider the following two topological frameworks.

To save space, this paper employs a one-dimensional linear notation to represent two-dimensional topological structures and evolutionary trees. For example: "(Root) ∠ Korean-Japanese = Turkic ≡ Mongolic ≒ Tungusic,"where the symbol "∠" represents an abstract root node symbolizing the starting point of evolution rather than any actual language. The symbol "-" indicates that the preceding language "Korean" represents the first (earliest) divergence; "=" indicates that "Japanese" is the second level of divergence; "≡" indicates that "Turkic" is the third level; and "≒" indicates that "Mongolic" is the fourth level, branching at the same level as the final language, "Tungusic."

As noted in the previous footnote, Indo-European was included in the experiment as an outgroup; here, it is treated as an ingroup merely to facilitate the understanding of linguistic relationships. Therefore, we place "Indo-European" in parentheses within the topological structure. Furthermore, if interpreted as an outgroup, the evolutionary structures generated by these two topologies after outgroup rooting are identical: ∠ Uralic-Turkic = Mongolic ≡ Tungusic.

Type A Topology: (Indo-European) = Uralic-Turkic = Mongolic ≡ Tungusic
Type B Topology: (Indo-European) - Uralic = Turkic ≡ Mongolic ≒ Tungusic

Type A treats Uralic and Indo-European as a (linguistic) alliance corresponding to Turkic, Mongolic, and Tungusic. Type B places Indo-European on one side, corresponding to the Uralic-Altaic macro-family on the other. The reason for these two types of structures is the aforementioned language count or the substitution of different individual languages within the set, a point that will be discussed in detail in the following sections and Section 3. These two linear topologies correspond to the following geometric topological graphs (dendrograms). Due to space constraints, clustering tree diagrams will generally not be listed hereafter.

Why does the number of languages cause differences in clustering? Previous research has offered preliminary explanations, suggesting that languages or dialects in boundary regions borrow features from opposing languages through contact [**?**]. The quantity of features borrowed by a single language is related to the depth of contact—for instance, the frequency, hierarchy, and number of lexical tokens. The multi-regional distribution and borrowing within the same language accumulate borrowing features, increasing surface similarity and smoothing over the differences between derived features (from evolution) and borrowed features. For example, if a feature exists in both the outgroup and the ingroup, it could be a symplesiomorphy or a borrowed form. This depends on whether the feature is ubiquitous within the ingroup or exists only in specific languages, which relates to the unbalanced linguistic distribution of borrowing. These factors are reflected in the number of asymmetric features between the outgroup and the ingroup. In short, the reality is far more complex than being "conditioned by language count." Behind the number of languages in the experiment lie the types and quantities of features borrowed through contact. A high number of features indicates diverse borrowing types across phonology, lexicon, and morphology, signaling deep linguistic contact. The difference between Type A and Type B topologies is that in Type B, Indo-European is independent of the ingroup hypothesis, fully conforming to the core principle of outgroup independence, whereas in Type A, the outgroup is embedded within the ingroup languages. There is a reason for this: we posit that Type A clustering must be caused by long-term proximity between Indo-European and Uralic, resulting in a large number of borrowed features in certain languages. Given that Type A clusters involve too many languages to list individually, they are omitted here.

We now expand the discussion on Type B clustering with specific examples. Through repeated experiments with specific languages and counts, we achieved Type B under the condition of 15 languages across 4 families. Type B implies that with Indo-European as the outgroup, the rooted tree shows that Uralic was the first to diverge from the Uralic-Altaic macro-family (see [Figure 3: see original paper]).

The "trees" initially generated by clustering algorithms are unrooted, showing only the relative relationships and branching nodes between taxa. Rooting provides the "tree" with a

temporal sequence. Rooting can be an internal

pre-set algorithmic rule.

To examine the factors affecting experimental results, the following experiments kept the Altaic languages constant while varying the number or identity of languages in the Uralic or Indo-European outgroups. For example, we replaced three Uralic languages in [Figure 3: see original paper] (Meadow Mari, Erzya, Komi > Hill Mari, Moksha, Udmurt). Although the sub-branches of these three replaced languages remained the same, the experimental result remained Type B. This is because both experiments involved peripheral minor languages with shallow contact with Indo-European.

Refer to [Figure 4: see original paper]. Furthermore, when we replaced Mari in the Uralic family (from [Figure 4: see original paper]) with Hungarian (Ugric branch) and Karelian (Finnic branch), the experimental result shifted to Type A (see [Figure 5: see original paper]). Why is this? Historical records show that linguistic contact between the Uralic and Indo-European families has a long history, extending from the Bronze Age (approx. 3300–2100 BCE) to the present. These changes are primarily manifested in major Uralic languages such as Hungarian and Finnish. For instance, the relationship between Hungarian and German is extremely close; the two were integrated historically through the Habsburg Monarchy and the Austro-Hungarian Empire (16th–20th centuries), leading to population mixing and typical language convergence. This experiment reflects the significant role played by borrowed linguistic features.

To verify the deeper reasons for classification differences caused by a language' s own borrowing features, we replaced the feature-rich Hungarian in [Figure 5: see original paper] with Northern Mansi from the same branch. As expected, the topology returned to Type B (see [Figure 6: see original paper]). This is because Northern Mansi has few loan features; it is an endangered language currently spoken by only about 2,000 people in the Siberian region of Russia. This phenomenon demonstrates that Uralic languages with complex features (like Hungarian) share more borrowed features with Indo-European and are thus prone to forming "alliances." Conversely, languages with fewer borrowed features have looser associations with Indo-European.

In the above experiments, the number of languages per family was generally 4–5, totaling approximately 15–20 languages for the four families (excluding the outgroup). This selection served two subjective purposes: first, a smaller number makes it easier to observe the causes of differences in linguistic relationships; second, this quantity keeps the system in a highly sensitive state where slight fluctuations can alter inter-family relationships, facilitating the adjustment of individual languages during experimentation. If we were to raise the system to a higher tier—approximately 7–8 languages per family, totaling 28–32 languages—the implications of adding or adjusting individual languages would be too broad and difficult to control. Through dozens of experiments involving the addition, removal, and replacement of individual Uralic, Turkic, and Indo-European outgroup languages, we found that at this higher magnitude, the system generally maintains a Type B structure. If the count is further increased to 10–12 lan-

guages per family (40–48 languages or more), the system primarily presents as Type A.

Overall, using Indo-European as an outgroup to verify the relationship between Uralic and Altaic is a choice of necessity. Aside from Indo-European, the only other options in the European and Eurasian border regions seem to be Caucasian languages. However, Caucasian languages—especially North Caucasian —contain a vast number of Turkic loanwords resulting from historical influxes of Turkic groups such as the Bulgars, Khazars, Kipchaks, and Oghuz, making them unsuitable as outgroups.

Through the adjustment of language counts and the replacement of individual languages with specific borrowing features, this section has preliminarily identified the quantitative conditions that reflect linguistic relationships under complex circumstances. The conclusion of this section is that the relationship between Uralic and Altaic languages is likely not one of common descent; their shared phonological, lexical, and morphological features are likely the result of mutual borrowing and convergence due to geographical proximity. This is precisely why they align either with the Indo-European outgroup or the Altaic group in experiments with different quantities and combinations. Therefore, through these quantitative comparative methods, the conclusions of this paper align with the mainstream academic view represented by Nicholas Poppe since the 1940s, which rejects the early Ural-Altaic hypothesis.

The Altaic Family and Radical Argumentation Methods

第二节有关乌拉尔语的讨论涉及一个重要概念，即参与实验的语言数量多寡会

This concept influences linguistic clustering relationships and can be referred to as the "language quantity parameter." It can be derived as a general theory and articulated through logical narration. We know that the calculation of linguistic similarity can be compared to the calculation of homologous evolution in species, where evolutionary distances and clustering can be used to construct phylogenetic trees [**?**, **?**, **?**]. However, even when relying on existing a priori knowledge, the languages selected for experimentation may include both cognate and non-cognate sets. This necessitates a process of discrimination and an exploration of the reasons why these factors influence the construction of phylogenetic trees.

Reviewing the experiments in the previous section, we observed that when the number of Turkic languages was increased, the topological relationship between Uralic and Indo-European languages could be interpreted as either nested or parallel (compare [Figure 2: see original paper] and [Figure 1: see original paper]). Once the technical procedure of Indo-European rooting was implemented, it became clear that Uralic was the first group to diverge from the so-called Ural-Altaic family. Indo-European is a known independent language family; therefore, its topological relationship with Uralic is by no means caused by genetic commonality. The only possible inference is that the two experienced close con-

tact and feature borrowing. In other words, since Indo-European, Uralic, and Altaic are generally considered to belong to different language families, the clusters they form can only be clusters of similar features resulting from geographical proximity and contact-induced borrowing. This includes Uralic borrowing Indo-European features, Turkic borrowing Uralic features, or vice versa. Because borrowing and lending are primarily realized by individual languages, the linguistic features borrowed into each language—and the quantity of those features—differ. This can lead to variations in Levenshtein distance during comparative calculations, thereby resulting in different clustering relationships. In summary, when non-genetically related languages are adjacent, one language may borrow features from another, causing the borrower and lender to share the same traits and potentially creating a similarity relationship. Furthermore, when the same language is distributed across different locations (i.e., dialects), each may borrow features from its respective neighboring languages. Consequently, the greater the number of languages involved in this borrowing and lending process, the more shared features there may be. This is the underlying reason why the number of selected languages impacts automatic clustering experiments.

Under normal circumstances, linguistic features shared through geographical proximity can be termed "shallow features," referring primarily to features borrowed over approximately one hundred years, or three to four generations. Their quantity fluctuates depending on socio-political factors; they float on the surface of the language and can sometimes be identified at a glance. These features may consist of new phonemes, new phonological structures, new vocabulary, morphological forms, or even unique phonetic and prosodic characteristics such as vowel harmony. However, relative to the source language, borrowed features tend to be isolated and fragmented; while they increase the variety of feature types, the number of instances within each type is relatively small. It must be noted that borrowed elements are not merely the product of a single time or place. From a longer historical perspective, another category can be established: "early features." These refer to linguistic features borrowed through contact over hundreds of years. Because borrowed phonetic forms evolve alongside the recipient language—where unique phonemes may either merge with original phonemes or be assimilated into the rules of the recipient language—some borrowed elements become formally quite close to the native features of the language. In short, these types of features stand in contrast to the "ontological features" of a language, generally referred to as cognate features, which are identical traits possessed by languages evolving from a common mother tongue. Compared to borrowed features, cognate features are fundamental and universal in both the phonetic system and the phonetic composition of words, as determined by the commonalities of human language.

Conversely, borrowed features may be unconventional or idiosyncratic; some languages possess them while others do not, and they are unevenly distributed across languages or dialects. Compared to shallow features, early features lack obvious foreign characteristics and are difficult to detect without systematic comparison. Regardless, the categories of early borrowed features should be

relatively few, and individual instances are unlikely to be numerous. Furthermore, while most loanwords occur in cultural vocabulary—and the nature of the Swadesh list used in these experiments limits the entry of borrowed features to some extent—they cannot be entirely eliminated. Early features, in particular, remain especially difficult to exclude.

3.2 偏激进实验目的和方案

We now turn our discussion to the Altaic hypothesis. Proponents argue that the Altaic family originates from structural linguistic commonalities shared among Turkic, Mongolic, and Tungusic languages, such as their agglutinative nature and Subject-Object-Verb (SOV) word order.

The alignment of characters calculated via Levenshtein distance is a specialized technical term referring to the process of assigning values to lexical differences and calculating the similarity distance between languages through the alignment of characters that constitute word pronunciations. See Jiang (2022).

Similarities in word order, vowel harmony, and suffixation, as well as shared vocabulary in personal pronouns—particularly the similar distribution and evolution patterns of certain consonants—are often cited. These shared features are theoretically regarded as manifestations of cognacy. Conversely, opponents argue that the commonalities within this language family are more likely the result of convergent evolution driven by long-term geographical proximity and linguistic contact rather than a common origin. Specifically, they point to a lack of sufficient cognates among Turkic, Mongolic, and Tungusic, noting that differences in core vocabulary, such as numerals, are significant across these groups. This discrepancy contradicts the expectations of the genetic relationship hypothesis within the framework of the comparative method. Proponents counter that the Altaic family is exceptionally ancient, with a divergence time so remote that cognates have been lost over time, potentially reaching a temporal depth where the comparative method is no longer applicable. Opponents emphasize that, according to historical records, early forms of Mongolic and Turkic often became more similar over time rather than diverging, as would be expected of a genetic relationship, thereby providing counter-evidence that these similarities result from language contact.

The Altaic hypothesis began with the concept of cognacy, but its argumentation encountered difficulties regarding contact-induced borrowing. The "Language Quantity Parameter" proposed above was originally a technical concept; however, methodologically, it can reveal relationships between languages through the depth of linguistic contact and the number of borrowed features. This serves as one of the primary argumentative methods relied upon in this paper. This section takes the opportunity to observe the relationships between the three language groups within the Altaic hypothesis through experimental observation. If these relationships can be determined using the aforementioned "Language Quantity Parameter," it may be possible to bypass the circular rea-

soning of the opposing sides. Furthermore, if the experimental results can be interpreted through major historical events involving the relevant regions and ethnic groups over the past millennium, this ancient proposition could receive a new interpretation.

According to the predictive experiments on the relationships between the three language groups in this paper, the Mongolic and Tungusic groups are more closely and stably integrated, while the Turkic group exhibits a degree of divergence from the former two. This paper designs a somewhat radical experimental scheme with a clear objective: to isolate the Altaic family by intentionally adjusting the number of languages in each group, thereby stripping the Turkic group away from the Altaic family. We then logically analyze the reasons for this separation or, conversely, analyze the reasons why it cannot be stripped away. In terms of specific operations, we randomly selected non-overlapping languages from the three Altaic groups to form two parallel test groups, Group A and Group B. Each group consists of 5 Turkic languages, 5 Mongolic languages, and 5 Tungusic languages, totaling 15 languages per group with no overlap between the two groups.

The experiment employs the outgroup method to determine linguistic relationships through outgroup rooting. Generally, the clustering before rooting is referred to as the topological structure, while the structure after rooting is called the evolutionary structure. An outgroup consists of one or more taxa included in the study that do not belong to the target languages (the ingroup). The function of the outgroup is to locate the root node of the evolutionary tree by introducing one or more taxa related to the ingroup but not part of it, thereby "orienting" the tree. Theoretically, the outgroup is assumed to share a more ancient common ancestor with the ingroup; thus, algorithmic techniques can place the root node at the intersection of the outgroup and ingroup branches. However, the selection of an outgroup is subject to certain constraints; improper selection can lead to errors, such as sequences becoming incomparable if the relationship is too distant, or the outgroup becoming nested within the ingroup if the relationship is too close.

3.3 数量调节：可监控的实验

The experiment was conducted in three stages. In the first stage, different outgroups were used to observe the topological and evolutionary structures of Group A and Group B. This stage also examined whether the outgroups exerted influence on or altered the internal structures of the three language families. According to the experimental design, the first stage utilized a near-minimum number of languages, with only five languages selected from each family. Based on the findings in Section 2, this represents a sensitive numerical state intended to test structural stability. To this end, outgroups were selected from different language families and geographical regions to ensure that genetic and contact similarities with the ingroup languages were minimized. Following this principle, we selected Basque, which is the only "clean" language isolate in Europe in terms

of its genealogical identity. We also selected Tamil, an ancient language of the Dravidian family in South Asia, and Chukchi, a Paleo-Siberian language of a reindeer-herding people in the Russian Far East. Ramstedt, in his *Introduction to Altaic Linguistics* (2004/1957), speculated through lexical comparison that the divergence of Proto-Altaic might have occurred earlier than that of the Indo-European family, the latter dating to approximately 6000–4000 BCE.

See Poppe, *Introduction to Altaic Linguistics*, pp. 230, 239, 240, whose views differ from those presented in this paper. Poppe points out that Turkic

and Mongolic exhibit systematic correspondences in vowel harmony rules and verbal suffixes, whereas the Tungusic verbal system is more complex and lacks similar regularities, suggesting that the former two may share a closer genetic relationship.

To ensure greater robustness, we also conducted experiments using multiple outgroup sets—that is, using several languages simultaneously as an outgroup. For this, we selected a combination of Nahuatl and Tarahumara from the Uto-Aztecan family of the Americas (Mexico). The experimental results consistently displayed the hypothesized internal boundaries of the Altaic family, all forming the same topological type: Turkic—Mongolic=Tungusic, hereafter referred to as Type C. Upon implementing rooting, the evolutionary structure remained unchanged, indicating that the internal relationships within the language family are quite stable. However, subtle variations appeared in two experiments where the Mongolic family split into two branches, resulting in a "Turkic—West Mongolic=East Mongolic Tungusic" topology (tentatively named Type Cx). This variation is related to the geographical distribution of the Mongolic family across thousands of kilometers between the East and West; the Eastern branch shows closer ties with Tungusic, though this does not fundamentally affect the overall analysis. Micro-analysis further reveals that Group B is more sensitive than Group A, which will be analyzed further below. Please refer to [Figure 7: see original paper]–10.

The second stage implemented a more radical scheme by selecting European languages that may have had extensive contact and borrowing with the Altaic family (primarily Turkic) as outgroups. The objective was to observe whether the outgroups could embed themselves into the ingroup and "pull" Turkic away from the original Altaic family to form a new alliance topology: [European Outgroup=Turkic]—[Mongolic=Tungusic]. This resembles the topological structure of Indo-European and Uralic discussed in Section 2. This argumentative strategy can be termed "bending" —an attempt to challenge established cognitive frameworks and propose alternative schemes through reverse thinking. The number of outgroup languages tested ranged from 2 to 12. Except for the occasional appearance of Type Cx in Group B, both Group A and Group B generally exhibited a Type C topology. The languages used were carefully selected, all having histories of long-term contact and fusion with Turkic and Mongolic. Examples include Russian, Polish, Ukrainian, and Albanian from the Indo-European family, and Hungarian and Udmurt from the Finno-Ugric

family. Further details are provided below.

The third stage involved further increasing the number of ingroup languages. On one hand, to further strengthen the relationship between European languages and Turkic, languages from the Uralic and Caucasian regions were added, bringing the number of outgroup languages to as many as 18:

eight Indo-European languages, six Uralic languages, and four Caucasian languages. On the other hand, since the number of outgroup languages far exceeded the number of languages in any single ingroup family—especially if one intends to strengthen the contact-based features between the outgroup and Turkic—increasing the number of Turkic languages was a viable approach. To this end, we merged the Turkic languages from Groups A and B and made minor adjustments to individual languages to create two new ingroup sets, Group Ab and Group Bb. These sets contained 11 Turkic languages, while the Mongolic and Tungusic families each maintained five languages. For the relevant experiments, see [Figure 11: see original paper] and [Figure 12: see original paper].

Strikingly, before rooting was implemented, Group Bb maintained the Type C structure ([Figure 11: see original paper]), while Group Ab appeared to fall into complete disarray. The Uralic languages were entirely embedded within the ingroup and allied with Turkic, severing the direct relationship with the Mongolic-Tungusic group. Furthermore, a portion of the Caucasian languages embedded themselves within the Tungusic languages, and the languages used as outgroups became fragmented and incomplete ([Figure 12: see original paper]). Why did this occur? Logically, this was likely caused by the numerical imbalance of languages between the families, which undoubtedly included feature-borrowing relationships between individual languages. This prompted us to merge the Mongolic and Tungusic languages from both groups, bringing the total number of ingroup languages to 31: 11 Turkic, 10 Mongolic, and 10 Tungusic.

The experimental results once again yielded a regular Type C topological structure (see [Figure 13: see original paper]). As a verification, we replaced all 10 Mongolic languages (Group Ac). The experimental results showed essentially no change in the internal topological structure; only the outgroup languages split into two branches: one for Indo-European and another for Uralic and Caucasian languages (see [Figure 14: see original paper]). If rooting is implemented, these merge into a single outgroup. This also demonstrates that stability is achieved when the number of internal and outgroup languages reaches a balance.

3.4 意外发现：蒙古语与满通古斯语的关系

The “three-step”clustering experiments described above clearly demonstrate that the Altaic hypothesis largely maintains a topological clustering result where the three language families remain connected, even under conditions where the number of languages and individual language samples are repeatedly varied. This suggests an inherent structural foundation, implying that a genetic relationship

may indeed exist among the three traditional Altaic groups. However, we cannot entirely overlook the disordered topological state observed in Group B of [Figure 12: see original paper]. This phenomenon must be driven by underlying factors, specifically the embedding of Caucasian languages within the Mongolic-Tungusic core group and the embedding of Uralic languages within the Turkic group. Therefore, even though this section adopts a relatively radical argumentative approach, it remains difficult to dismiss the Altaic hypothesis outright. One point, however, is relatively clear: the Mongolic and Tungusic families form an exceptionally stable cluster. This stability perhaps justifies the hypothesis of a more probable genetic relationship: a "Mongolic-Tungusic" language family.

Given the limited corpus currently available, it is impossible to continue conducting experiments under the theme of "quantity." Consequently, the methodology employed in this paper cannot provide a definitive conclusion regarding the credibility of the Altaic hypothesis.

Clustering Experiments on Transeurasian Languages: Korean and Japanese

4.1 朝语系和日语系的独立实验

This section conducts a clustering study using five language groups within the Transeurasian family (Turkic, Mongolic, Tungusic, Koreanic, and Japonic) as experimental subjects. During implementation, two primary difficulties were encountered: first, how to determine the experimental outgroups; and second, how to determine whether the Koreanic and Japonic families exhibit a phylogenetically significant clustering with the Altaic family.

Regarding the first point, an outgroup is defined relative to an ingroup. It consists of one or more taxonomic units related to the ingroup languages, serving to root the "tree" and helping to determine the evolutionary relationships of other units and the position of the "root." Theoretically, an outgroup can be set as a "sister group" to the ingroup, with both sharing a distant common ancestor.

In this sense, geographic proximity is a significant indicator of a "sister group." If Koreanic is considered a member of the Altaic family, there are essentially only two directions for selecting an outgroup geographically. To the north are the Northeast Asian (NEA) languages, also known as Paleosiberian languages, such as Ket (Yeniseian), Chukchi (Chukotko-Kamchatkan), and languages of uncertain affiliation like Yukaghir and Nivkh. To the south are the Sinitic (Chinese) dialects. The former are distributed across the Russian Far East and Siberia, extending to the lower reaches of the Heilongjiang River (referred to as the Amur River in Russia) and northern Sakhalin Island—specifically the present-day Khabarovsk Krai and Sakhalin Oblast of the Russian Federation. Historically, the Japonic family was isolated on an archipelago; compared to the languages of mainland East Asia, it possesses unique characteristics in both phonological structure and morphological type, and is generally regarded as an independent language or language family. This paper argues that because

Koreanic and Japonic possess entirely different properties, it is difficult to determine a single experimental outgroup for them when they are included in a large macro-family. Therefore, this experiment attempts to find separate outgroups for each to identify underlying patterns before exploring a possible common outgroup. According to the principle of geographic proximity, there appear to be only two possibilities for Japonic outgroups.

The first is Austronesian (AN), including languages from Taiwan and Southeast Asian countries to the south, or those from the broader East and Southeast regions of the South Pacific. The second is Sinitic (Chinese dialects). The Austronesian outgroups selected for this study are divided into two groups: Group P includes Hawaiian, Fijian, and Tahitian from the East and Southeast Pacific; Group Q includes Amis, Tongan, and Tagalog from the Southwest Pacific. The Sinitic dialects used are Jinan, Taiyuan, and Xi' an, which also serve as the outgroups for the Koreanic experiments discussed below.

Regarding the second point, this question must be answered through a series of experiments. The linguistic data used in this paper adds Japonic and Koreanic languages to the Altaic groups established in Section 3. While Koreanic is generally considered a primary language family, some scholars argue that the Jeju dialect of Korean can be classified as an independent language. This experiment selects five languages or dialects from the Koreanic family: the Pyongyang, Seoul, Jeolla, and Jeju dialects, as well as the Yanbian Korean dialect distributed in Jilin Province, China. As a potential language family, Japonic is generally considered to include Ainu (Hokkaido and Sakhalin) and Ryukyuan (Miyako and Amami). In addition to collecting four dialects from these two languages, this experiment also selects several Japanese dialects, including Tokyo, Kyoto, Fukuoka, Hachijo, Hokkaido, and Kagoshima; see the experimental data preview in Section 2.

Following the experimental protocols of the previous two sections, we similarly set up groups of five languages, further divided into Groups A and B. However, due to the limited number of Koreanic dialects collected, each experiment must share the same set of languages. The ten Japonic languages or dialects are exactly sufficient to be allocated into two groups for the experiments.

Historically, the concept of a "Koreo-Japonic" family was proposed, primarily during the period of Japanese colonization of Korea (1910–1945).

For example, Kanazawa Shōzaburō proposed the "Theory of the Common Origin of Japanese and Korean" (1929). Additionally, some contemporary historians and government officials held similar views, which were driven by the political requirements of colonial rule rather than scientific motivation.

Experiment 1: (Altaic + Koreanic + Northeast Asian outgroup (NEA)), Groups A and B. The results are: Topology: Turkic(+NEA)-Korean=Mongolic Tungusic; Rooting: Turkic-Korean=Mongolic Tungusic. This experiment reveals a connection between the outgroup languages and the Turkic languages within the ingroup; the NEA outgroup is embedded within Turkic and shows no strong

relationship with Koreanic. According to Wikipedia and other sources, the so-called Northeast Asian languages of the Russian Far East, Kamchatka Peninsula, Chukchi Peninsula, and Sakhalin Island are usually regarded as remnants of ancient indigenous languages that do not belong to any known major language family. However, geographically adjacent to or even overlapping with these Paleosiberian languages is Sakha (Yakut), a Turkic language that expanded into the region (Sakha Republic, Russian Far East). This is a primary reason why the NEA outgroup embeds within the Turkic group, suggesting a relationship based on borrowing. See Figure 15 [Figure 15: see original paper].

Experiment 2: (Altaic + Koreanic + Sinitic outgroup (CH)), Groups A and B. The results are: Topology: Turkic=Korean(+CH)-Mongolic=Tungusic; Rooting: Korean-Turkic=Mongolic Tungusic. The results for the Altaic and Koreanic experiments with Sinitic outgroups are consistent across Groups A and B, as shown in Figures 16a [Figure 16: see original paper] and 16b.

Northeast Asian Outgroup - Topological/Evolutionary Structure (((Sinitic Outgroup - Topological Structure (((((Sinitic Outgroup - Evolutionary Structure

The topological structure of Experiment 2a directly demonstrates a close correlation between Koreanic and the Sinitic outgroup, which perfectly aligns with the historical fact of heavy Sinitic influence on Koreanic. However, the rooting results for Koreanic under different outgroup conditions are puzzling. From the perspective of rooting (Figure 16b), Experiment 2b shows Koreanic diverging before Turkic, whereas Experiment 1 shows Turkic diverging first and Koreanic later. Such inconsistent results make it difficult to claim that Koreanic is a stable member of the Altaic family.

Experiment 3: (Altaic + Japonic + Austronesian outgroup (AN)), further divided into Austronesian outgroup groups P and Q. Outgroup P, Groups A and B topology: Japonic(+AN)=Turkic-Mongolic=Tungusic; Rooting: Japonic-Turkic=Mongolic Tungusic. Outgroup Q, Group A topology: Japonic(+AN)=Turkic-Mongolic=Tungusic; Rooting: Japonic-Turkic=Mongolic Tungusic. Outgroup Q, Group B topology: Japonic=Turkic(+AN)-Mongolic=Tungusic; Rooting: Turkic-Japonic=Mongolic Tungusic. We note that the topological and rooting structures of Group A for outgroup Q are consistent with both groups for outgroup P: the Austronesian outgroup attaches to Japonic, implying feature borrowing between Japonic and Austronesian (Benedict 1990, Sagart 2011). However, under the condition of outgroup Q, Group B undergoes a significant change. The reason may be that the Austronesian languages in outgroup Q share similarities with the Turkic languages in Group B due to borrowing or other factors, the exact nature of which we have not yet determined. Nevertheless, this phenomenon is not an isolated case; it was also clear in Experiments 1 and 2 regarding the relationship between Koreanic and Altaic: the NEA outgroup has a contact-borrowing relationship with Turkic, as discussed in detail by Piispanen (2013). Conversely, the Sinitic outgroup reveals the frequent historical contact between Koreanic and Sinitic.

Experiment 4: (Altaic + Japonic + Sinitic outgroup (CH)). The results are consistent with the Austronesian outgroup (P) experiment. The four experiments above regarding Koreanic and Japonic all involve contact relationships between the outgroup and ingroup languages, which is detrimental to judging the phylogenetic relationship between Koreanic or Japonic and the Altaic languages. To eliminate the influence of contact borrowing, we attempted to select ultra-distant outgroup languages. "Ultra-distant" refers to both great geographical distance and different genealogical classifications. The following experiments use the Uto-Aztecan family of the Americas and the Dravidian family of South Asia as outgroups to observe the relationship between Koreanic or Japonic and Altaic.

Experiment 5: (Altaic + Koreanic + Uto-Aztecan outgroup (UT)), Groups A and B. The results are: Group A: Topology: Korean(+UT)=Turkic-Mongolic=Tungusic. Group B: Topology: Turkic-Korean(+UT)=Mongolic Tungusic; Rooting: Korean-Turkic=Mongolic Tungusic. Experiment 6: (Altaic + Koreanic + Dravidian outgroup (DR)), Groups A and B. The results are:

Topology: Korean(+DR)=Turkic-Mongolic=Tungusic; Rooting: Korean-Turkic=Mongolic Tungusic. Experiment 7: (Altaic + Japonic + Uto-Aztecan outgroup (UT)), Groups A and B. The results are:

Group A: Topology: Japonic(+UT)-Turkic=Mongolic Tungusic. Group B: Topology: Japonic(+UT)=Turkic-Mongolic=Tungusic; Rooting: Japonic-Turkic=Mongolic Tungusic. Experiment 8: (Altaic + Japonic + Dravidian outgroup (DR)), Groups A and B. The results are:

Topology: Japonic(+DR)=Turkic-Mongolic=Tungusic; Rooting: Japonic-Turkic=Mongolic Tungusic. The purpose of these four sets of experiments is to exclude borrowing-related associations between Koreanic or Japonic and the outgroup languages. The fact that the outgroup attaches to Koreanic or Japonic without being embedded within the ingroup is likely due to computational processing techniques and does not represent a necessary connection between them and the ingroup. However, the rooting consistently forces Koreanic or Japonic to isolate from the Altaic family first. These ambiguous results make it difficult to determine whether a cognate relationship exists between Koreanic/Japonic and the Altaic family. Thus, this line of argumentation appears to be a dead end.

However, we might make a further logical hypothesis: if Koreanic and Japonic have no phylogenetic relationship with the Altaic family, what would happen if we replaced them with another language family known to be unrelated to Altaic, such as Sino-Tibetan, and applied the same experimental framework?

Experiment 9: (Altaic + Sino-Tibetan + Uto-Aztecan outgroup (UT)), Groups A and B. The results are: Topology: Sino-Tibetan(+UT)-Turkic=Mongolic Tungusic; Rooting: Sino-Tibetan-Turkic=Mongolic Tungusic. Experiment 10: (Altaic + Sino-Tibetan + Dravidian outgroup (DR)), Groups A and B. The results are:

Group A: Topology: Sino-Tibetan(+DR)-Turkic=Mongolic Tungusic. Group B: Sino-Tibetan-(+DR)=Turkic Mongolic Tungusic; Sino-Tibetan-Turkic=Mongolic Tungusic. Sino-Tibetan has no phylogenetic relationship with the Altaic languages. The experimental results show that, except for a slight difference in the topological structure of Experiment 10 (where the outgroup is independent and not attached), the other structures—especially the evolutionary structures—are no different from those of Koreanic and Japonic described above. Since it is known that Sino-Tibetan is unrelated to Altaic, and the experiment merely replaced Koreanic or Japonic while keeping the conditions, methods, language selection, and grouping identical, this study logically confirms that neither Koreanic nor Japonic belongs to the Altaic family.

4.2 朝语系和日语系的合并实验

In addition to the strategy of alternating outgroups to test Koreanic and Japonic languages individually—or directly replacing them with Sino-Tibetan languages—another viable strategy involves simultaneously including both Koreanic and Japonic languages in the test to observe their clustering behavior within the experiment. As previously established, the outgroup languages for this experiment must be ultra-distant. Consequently, we continue to employ the Uto-Aztecan languages of North America and the Dravidian languages of South Asia as our designated outgroups.

Experiment 11: Phylogenetic Analysis of Altaic, Koreanic, Japonic, and Uto-Aztecan (UT) Outgroups

Experiment 11a: Group A Results

In Experiment 11a, we conducted a phylogenetic analysis involving the Altaic languages, Koreanic, Japonic, and the Uto-Aztecan (UT) family as an outgroup. The experimental results for Group A are as follows:

Topology: The recovered topology indicates a primary grouping of Japonic and Koreanic relative to the outgroup (UT), followed by a clade consisting of Turkic, Mongolic, and Tungusic, represented as: Japonic = Koreanic (UT) - Turkic = Mongolic Tungusic.

Rooting: The rooted tree structure follows the pattern: Koreanic - Japonic = Turkic Mongolic Tungusic.

Experiment 11b: Group B Results

Experiment 11b utilized the same language families—Altaic, Koreanic, Japonic, and the Uto-Aztecan (UT) outgroup—but focused on the Group B dataset. The experimental results for Group B are as follows:

Results and Analysis

Topology and Rooting Patterns

The experimental results reveal a specific topological structure: Japanese ≡ Korean (UT) = Turkic-Mongolic = Tungusic. When examining the rooting, we observe the pattern: ∠ Korean-Japanese = Turkic ≡ Mongolic ≈ Tungusic.

In Experiment 12, which included Altaic, Koreanic, Japonic, and a Dravidian outgroup, two distinct configurations (Group A and Group B) emerged. The topology for Group B was identified as: Japanese (DR) ≡ Korean = Turkic-Mongolic = Tungusic, with a corresponding rooting of ∠ Japanese-Korean = Turkic ≡ Mongolic ≈ Tungusic (DR).

Discussion on Phylogenetic Stability

A critical observation from these experiments is that alternating the outgroup triggers significant changes in the topological structure. Specifically, the evolutionary structure of the phylogenetic relationships—represented by the rooting which characterizes the sequence of evolution—undergoes shifts. In a real-world evolutionary context, such instability is impossible.

The primary objective of these experiments is to uncover the underlying historical truth of linguistic divergence. If the Koreanic and Japonic language families were indeed genuine members of a "Transeurasian" macro-family, their relative order of divergence should remain constant regardless of the outgroup used. The phenomenon observed in these experiments—where Koreanic and Japonic alternate in their sequence of differentiation—is an unacceptable result that contradicts the principles of stable phylogenetic descent.

The genetic relationship between the Koreanic and Japonic language families and the Turkic, Mongolic, and Tungusic families was primarily systematized and articulated by Ramstedt (2004/1957). He provided extensive evidence by reconstructing the roots and morphological features of Proto-Altaic, utilizing methods centered on phonetic correspondence and reconstruction. Compared to these early historical comparative methods, the quantitative comparison employed in this study allows for further breakthroughs. For instance, by moving beyond the frameworks established by previous scholars and employing permutation clustering of languages or language families, we can derive more objective conclusions from the clustering results. Logically speaking, the findings in Section 4.1 already provide a negative conclusion regarding the individual entry of the Japonic and Koreanic families into the Altaic phylum; consequently, their simultaneous inclusion in the Altaic family is even more improbable. However, considering the inherent randomness of evolution, this section will nonetheless proceed with the relevant experiments.

This approach mirrors the methodology previously employed, where the Sino-Tibetan language family was used to substitute for the Koreanic or Japonic families. If the experimental results of merging the Koreanic and Japonic families

are consistent with the individual results presented in Section 4.1, it would further indicate that neither the Koreanic nor the Japonic language family shares a genetic relationship with the Altaic family. To ensure a more robust demonstration of these findings, the Austronesian language family could also be used to replace the Koreanic and Japonic families in subsequent experiments.

2.2 Austronesian Language Experiments

The Austronesian languages are similarly divided into Group A and Group B for experimental purposes. Following the methodology established in previous sections, the outgroup languages consist of Uto-Aztecan languages from South America and Dravidian languages from South Asia. The following sections detail the results of these grouped experiments.

Experiment Results and Phylogenetic Analysis

Experiment 13: Altaic, Koreanic, and Austronesian with Uto-Aztecan (UT) Outgroup

In Experiment 13, we analyzed the linguistic relationships between the Altaic, Koreanic, and Austronesian families using the Uto-Aztecan (UT) family as an outgroup. For both Group A and Group B, the resulting topology is represented as: Koreanic (+UT) = Austronesian - Turkic = Mongolic Tungusic. The rooted tree structure follows the pattern: Koreanic - Austronesian = Turkic Mongolic Tungusic. These results suggest a specific branching order where Koreanic and the outgroup show a closer proximity relative to the broader Altaic-Austronesian cluster.

Experiment 14: Altaic, Koreanic, and Austronesian with Dravidian (DR) Outgroup

Experiment 14 utilized the Dravidian (DR) family as the outgroup to examine the same core language families. For both Group A and Group B, the topology is identified as: [Koreanic = Austronesian] (+DR) - Turkic = Mongolic Tungusic. The rooted configuration is expressed as: [Koreanic = Austronesian] - Turkic = Mongolic Tungusic. This indicates a stable grouping of Koreanic and Austronesian when contrasted against the Dravidian outgroup and the remaining Altaic branches.

Experiment 15: Altaic, Japonic, and Austronesian with Uto-Aztecan (UT) Outgroup

In Experiment 15, the Koreanic family was replaced by the Japonic family. Using the Uto-Aztecan (UT) outgroup, the topology for both Group A and Group B was determined to be: Japonic (+UT) = Austronesian - Turkic = Mongolic Tungusic. The rooted tree structure is: Japonic - Austronesian = Turkic Mongolic Tungusic. This topology mirrors the results found in

Experiment 13, substituting Japonic for Koreanic in the primary branching position relative to the outgroup.

Experiment 16: Altaic, Japonic, and Austronesian with Dravidian (DR) Outgroup

Experiment 16 investigated the relationships between the Altaic, Japonic, and Austronesian families using the Dravidian (DR) family as an outgroup. The

Group A: Topology: Japanese=Austronesian-Turkic=Mongolic Tungusic; Rooting: [Japanese=Austronesian]-Turkic=Mongolic Tungusic. Group B: Topology: Japanese(+DR)=Austronesian-Turkic=Mongolic Tungusic; Rooting: Japanese-Austronesian=Turkic Mongolic Tungusic. Experiments 13 and 15 indicate that after outgroup rooting, the first to differentiate are the Koreanic or Japonic families, occurring prior to the divergence of the Austronesian languages. This result carries only two possible implications: either the Koreanic/Japonic families and the Austronesian languages bear no relationship to the three language families of Altaic, or the relationship between Austronesian and Altaic is closer than that of Koreanic and Japonic to Altaic.

Furthermore, examining the rooting results of Experiment 14 (Groups A and B) and Experiment 16 (Group A), we observe that under Dravidian outgroup conditions, Koreanic and Austronesian (or Japonic and Austronesian) form a cluster that differentiates first. However, while keeping the outgroup constant, the results of Experiment 16 (Group B) differ from those of Group A. We can only speculate that various languages may carry "impurities" that differ entirely in both nature and quantity. These so-called impurities consist of various types of external loan features, which create scattered shared similarities. The consequence of these features is the smoothing over of cognate differences between languages, making it extremely difficult for computational models to polarize homologous features.

Conclusion: The Mongolic-Manchu Hypothesis

The genetic relationship between the Mongolic and Manchu-Tungusic languages has long been a subject of intense debate within the field of Altaic linguistics. This study, through a systematic comparative analysis of core vocabulary, morphological structures, and phonological correspondences, re-evaluates the "Mongolic-Manchu Hypothesis." Our findings suggest that the similarities observed between these two language families extend beyond mere lexical borrowing resulting from long-term geographical proximity and cultural contact.

The evidence presented in this paper highlights a significant layer of shared grammatical markers and basic lexical items that are resistant to borrowing. Specifically, the correspondence patterns in the verbal inflection systems and the pronominal paradigms provide compelling support for a common ancestral stage. While the broader "Altaic" theory remains controversial, the closer relationship between Mongolic and Manchu-Tungusic—often referred to as the

"Eastern Altaic" or "Macro-Mongolic" group—appears increasingly robust when subjected to rigorous historical-comparative methods.

In conclusion, while we acknowledge the profound influence of historical prestige languages and the complex "Sprachbund" dynamics of Northeast Asia, the structural parallels identified in this research point toward a genetic affinity. Future studies utilizing computational phylogenetics and ancient DNA analysis will be essential to further refine the timeline of divergence and to distinguish between inherited traits and those acquired through the extensive horizontal transfer characteristic of the Inner Asian linguistic landscape.

Can the Transeurasian language family (macrofamily) truly be trusted? Specifically, regarding whether the Koreanic and Japonic languages can be incorporated into it, the arguments presented in Section 4 lead to a negative conclusion. Furthermore, can the traditional or classical Altaic hypothesis be sustained?

Due to limitations in the available corpora, Section 3 can only state that a definitive conclusion cannot be reached. Regarding the earlier Ural-Altaic hypothesis, Section 2 has already refuted it using quantitative comparative experimental methods. However, following this extensive series of arguments and discussions—corresponding to the three levels of genetic classification—the algorithmic model proposed in this paper suggests the existence of a stable language group. In linguistic terms, this would correspond to a "Mongolic-Tungusic" family. Regarding this potential phylogenetic hypothesis, we look forward to further discussion and verification by the academic community in the future.

Postscript: As this article was being completed, I happened to read a newly published paper by Martine Robbeets and Mark Hudson [**?**]. This study utilizes a Bayesian phylogenetic approach to analyze the dispersal of Transeurasian languages (including Japanese, Korean, Tungusic, Mongolic, and Turkic). The authors argue that the expansion of these language families was driven by the spread of millet agriculture from Northeast China, specifically the West Liao River basin, beginning in the Early Neolithic.

This "farming-language dispersal" hypothesis aligns closely with the archaeological evidence discussed in this paper regarding the northward transmission of agricultural techniques. Robbeets and Hudson suggest that the initial expansion of Transeurasian speakers was linked to the Hongshan culture and its successors, which matches our observation of the significant cultural influence exerted by the West Liao River region on the Russian Far East and the Korean Peninsula. Furthermore, their genetic analysis indicates a shared "Amur-like" ancestry among early Transeurasian speakers, providing independent biological support for the cultural and linguistic connections we have identified through the material record.

The integration of linguistic, genetic, and archaeological data in their work offers a compelling multidisciplinary framework that reinforces our conclusions. It underscores the pivotal role of the West Liao River basin as a primary center of innovation and a source for the demographic and cultural shifts that reshaped

the human landscape of Northeast Asia. These findings further validate the necessity of viewing the prehistoric development of the region through a broad, trans-regional lens that accounts for the complex interplay between environmental adaptation, technological diffusion, and human migration.

Introduction

The newly published volume, *The Oxford Handbook of Archaeology and Anthropology*, represents a significant milestone in the interdisciplinary study of human history. This comprehensive handbook brings together leading scholars to explore the complex intersections between material culture, linguistic evolution, and social structures. By integrating diverse methodological approaches, the work provides a robust framework for understanding how ancient societies communicated, organized themselves, and interacted with their environments.

Theoretical Frameworks and Methodologies

The integration of archaeology and linguistics requires a sophisticated theoretical foundation to bridge the gap between physical artifacts and intangible heritage. This section examines the evolution of "linguistic archaeology" and the application of comparative methods to reconstruct proto-languages alongside archaeological horizons.

1.1 Integrating Material and Linguistic Evidence

One of the primary challenges in this field is the synchronization of chronological data from disparate sources. While archaeology relies on stratigraphic sequences and radiocarbon dating, historical linguistics utilizes glottochronology and the comparative method. The handbook proposes a unified model where:

$$\mathcal{R} = \{A \cap L \mid T_a \approx T_l\}$$

In this expression, $\mathcal{R}$ represents the set of correlations where archaeological evidence A and linguistic evidence L align within a similar temporal window T. By identifying these points of convergence, researchers can more accurately map the migration patterns of early human populations.

1.2 Computational Approaches in Historical Linguistics

Recent advancements in machine learning and Bayesian phylogenetics have revolutionized our ability to model language dispersals. As noted in [**?**], these computational tools allow for the testing of competing hypotheses regarding the origins of major language families, such as Indo-European or Austronesian. The application of these models often involves complex probability distributions:

$$P(T \mid D) = \frac{P(D \mid T)P(T)}{P(D)}$$

where T is the phylogenetic tree and D is the observed linguistic data. This Bayesian approach provides a statistical basis for linking linguistic divergence with archaeological shifts in material culture.

[Figure 1: see original paper]

Case Studies in Regional Prehistory

The handbook provides detailed regional analyses that demonstrate the practical application of these interdisciplinary methods. From the spread of agriculture in the Fertile Crescent to the maritime expansions in the Pacific, the synthesis of data offers new insights into human mobility.

2.1 The Indo-European Expansion

The "Steppe Hypothesis" remains

Language, June 2025). According to a review by Juha Janhunen, a leading authority in Altaic studies, this marks the first time in over a quarter of a century that Oxford University Press has released such a significant work in this field.

A New Handbook of Archaeology and Linguistics: Integrating Archaeology, Genetics, and Linguistics

The primary objective of this handbook is to explore the methodological and theoretical frameworks for integrating archaeology, genetics, and linguistics. By synthesizing these three distinct yet complementary disciplines, we aim to develop a more comprehensive understanding of human history, population movements, and cultural evolution.

1.1 Interdisciplinary Synthesis

The integration of archaeology, linguistics, and ancient DNA (aDNA) research represents a transformative shift in the study of the human past. While archaeology provides the material context of human activity and linguistics traces the evolution and spread of languages, genetics offers direct insights into biological ancestry and migration patterns. This handbook addresses the challenges of aligning these different datasets, which often operate on different temporal and spatial scales.

1.2 Methodological Frameworks

To achieve a successful synthesis, it is essential to establish rigorous methodological standards. This involves:

- **Spatial and Temporal Correlation:** Developing models that can map linguistic phylogenies onto archaeological horizons and genetic clusters.
- **Quantitative Modeling:** Utilizing machine learning and statistical tools to analyze large-scale datasets across disciplines.
- **Terminology Standardization:** Ensuring that terms such as "culture," "population," and "language family" are used with precision to avoid circular reasoning.

1.3 Case Studies in Human Prehistory

The handbook examines several key historical transitions where the intersection of these fields has proven particularly fruitful. For instance, the expansion of the Indo-European language family has been re-evaluated through the lens of Yamnaya migrations identified in genomic studies, coupled with the archaeological evidence of the Bronze Age. Similarly, the dispersal of Austronesian languages and the spread of agriculture in Neolithic Europe serve as critical benchmarks for testing interdisciplinary hypotheses.

1.4 Future Directions

Looking forward, the field must move beyond simple correlations toward more nuanced models of interaction. This includes accounting for processes such as language shift without significant genetic replacement, or cultural diffusion that occurs independently of large-scale migration. By fostering a collaborative environment, we can move closer to a holistic "science of the past" that respects the unique contributions of each discipline while striving for a unified narrative of human heritage.

This handbook integrates archaeology, genetics, and linguistics into a comprehensive approach aimed at elucidating the human experience of the past and its potential trajectories for the future. The manual is divided into three parts: "Basic Framework: Archaeology, Genetics, and Linguistics," the second···

Crossing Time: Archaeology, Genetics, and Language

The study of human history and evolution has entered a new era characterized by the integration of diverse scientific disciplines. By synthesizing evidence from archaeology, genetics, and linguistics, researchers can reconstruct the complex migration patterns, cultural shifts, and biological adaptations that have shaped modern populations. This interdisciplinary approach allows for a more nuanced understanding of how ancient societies interacted and how their legacies persist in the contemporary world.

Archaeological Perspectives

Archaeology provides the material foundation for understanding past human behavior. Through the systematic excavation and analysis of artifacts, architecture, and environmental remains, archaeologists reconstruct the socio-economic structures and technological innovations of ancient civilizations. Recent advancements in dating techniques and remote sensing have significantly refined our chronological frameworks, enabling a more precise mapping of cultural transitions across different regions and epochs.

[Figure 1: see original paper]

Genetic Insights

The field of paleogenomics has revolutionized our understanding of human ancestry. By sequencing ancient DNA (aDNA) from skeletal remains, scientists can trace the genetic lineages of extinct populations and identify instances of admixture between different groups. These genetic data often reveal migrations and population replacements that are not always visible in the archaeological record alone. Furthermore, comparing ancient genomes with those of modern populations helps clarify the evolutionary pressures that have influenced human health and physiology over millennia.

Linguistic Evolution

Linguistics offers a unique window into the cognitive and social history of humanity. The comparative method in historical linguistics allows researchers to reconstruct ancestral languages (proto-languages) and trace the diversification of language families. When combined with archaeological and genetic data, linguistic patterns can provide critical clues regarding the spread of agricultural practices, the movement of nomadic tribes, and the formation of early states. The correlation between linguistic boundaries and genetic clusters often highlights long-term barriers to gene flow or, conversely, periods of intense cultural exchange.

Synthesis and Integration

The integration of these three fields—archaeology, genetics, and linguistics—creates a robust framework for historical synthesis. For instance, the expansion of specific language families can often be linked to the dispersal of genetic markers and the appearance of distinct archaeological assemblages. By cross-referencing these independent lines of evidence, scholars can develop more comprehensive models of human prehistory, moving beyond single-discipline narratives to a holistic view of our shared past. This "consilience" of evidence is essential for addressing fundamental questions about human origins, identity, and the enduring impact of ancient migrations on the modern world.

Paleolinguistics is applied across various stages of human history, spanning from

hunter-gatherer societies through the adoption of agriculture and the rise of writing, up to the modern era. Section 3···

Crossing Space: Archaeology, Genetics, and Linguistics

This section explores the intersection of archaeology, genetics, and linguistics, utilizing case studies from various regions across the globe to demonstrate the integrated application of "archaeological linguistics." By synthesizing these diverse disciplines, we can reconstruct the complex history of human migration, cultural evolution, and the dispersal of language families with unprecedented precision.

Global Case Studies in Interdisciplinary Research

The integration of ancient DNA (aDNA) analysis with traditional archaeological findings and historical linguistics has revolutionized our understanding of human prehistory. For instance, the expansion of the Indo-European language family has long been a subject of debate. Recent genomic evidence, when mapped against archaeological horizons such as the Yamnaya culture, provides a clearer picture of how both genes and languages spread across the Eurasian steppe into Europe and South Asia.

Similarly, in the Asia-Pacific region, the dispersal of Austronesian languages offers a compelling example of "archaeological linguistics"in action. By correlating the distribution of specific pottery styles (such as Lapita ware) with linguistic reconstructions of maritime technology and genetic markers found in modern and ancient populations, researchers can trace the "Out of Taiwan" migration route across the Pacific islands.

Methodological Integration

The synergy between these fields allows for a multi-dimensional approach to the past:

- **Archaeology** provides the material context, offering physical evidence of settlement patterns, subsistence strategies, and social structures.
- **Genetics** tracks the movement of biological populations, identifying instances of migration, admixture, and demographic shifts that may or may not align with cultural changes.
- **Linguistics** offers insights into the cognitive and social frameworks of ancient peoples, tracing the evolution of vocabulary and grammar to infer historical contacts and common ancestries.

By bridging these spatial and temporal gaps, researchers can move beyond isolated data points to construct a holistic narrative of human development. These case studies underscore the necessity of a transdisciplinary framework—where the limitations of one field are addressed by the strengths of another—to solve the enduring mysteries of our collective history.

This book explores the applications of linguistics across various disciplines. Comprising 33 chapters, the volume features contributions from 68 authors, including numerous professors from China. The work introduces a significant number of new terms and conceptual frameworks, such as "archaeolinguistics," "linguistic palaeontology," "linguistic phylogeography," "phylogenetic linguistics," and "pedigree linguistics."

(phylolinguistics), as well as several interdisciplinary technical neologisms and terms. At the time of this article' s publication, we include this latest development here for the reader' s reference. We also wish to quote once more from Hu Ning' s book review: "The 'triangulation' method advocated by Robbeets for exploring the history of languages and language families is easily criticized for a simple reason: genes and archaeological cultures do not speak. The key role in searching for the origins of language still belongs to historical linguistics." From this, it is evident that linguistics within interdisciplinary research still bears a heavy responsibility at the levels of human history and society.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ···& Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ···& Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[7] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[8] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

[10] Rumelhart, D. E., Hinton, G. E., & Williams, R. J.

Poppe, Nicholas. Translated by Zhou Jianqi, 2004/1965. *Comparative Grammar of the Altaic Languages* [阿尔泰语比较语法]. Hohhot: Inner Mongolia Education Press.

Chen Zongzhen et al., 1990, *Zhongguo Tujue Yuzu Yuyan Cihui Ji* [A Lexicon of the Turkic Languages of China]. Beijing: Minzu Chubanshe [Nationalities Publishing House].

Ethno Publishing House. Frye, R. N. Translated by Han Zhongyi, Fu Jiajie, and Shui Minjun, 2024. *The Heritage of Central Asia* [中亚古代史]. Shanghai: Ethno Publishing House.

Shanghai People' s Publishing House.

Automatic Clustering and Division of Chinese Dialects and Related Computational Methods

Abstract

The quantitative classification and regional division of Chinese dialects represent a core challenge in Chinese dialectology. Traditional qualitative methods, while foundational, often face difficulties when processing large-scale linguistic data or resolving ambiguous boundaries. This paper explores the application of automatic clustering algorithms and related computational methods to the classification of Chinese dialects. By integrating machine learning techniques with traditional dialectological theories, we propose a systematic framework for dialect division. We analyze the effectiveness of various distance metrics and clustering algorithms, such as K-means and hierarchical clustering, in capturing the phonetic, lexical, and grammatical variations across different dialect regions. The results demonstrate that computational methods can provide objective, reproducible, and fine-grained insights into the internal structure of Chinese dialects, offering a valuable complement to traditional qualitative analysis.

1. Introduction

The classification and division of Chinese dialects have long been a focal point of linguistic research. Since the pioneering work of early dialectologists, the identification of dialect boundaries has relied heavily on specific diagnostic features, such as the evolution of the Middle Chinese voiced initials and the distribution of tone categories. However, as the volume of dialectal data increases—particularly with the development of large-scale linguistic atlases and databases—manual classification becomes increasingly complex.

The introduction of computational methods offers a path toward more rigorous and objective dialectometry. By treating dialectal variations as multidimensional data points, researchers can employ clustering algorithms to identify natural groupings within the data. This paper aims to evaluate the current

state of automatic clustering in Chinese dialectology and propose optimized computational strategies for dialect division.

2. Computational Framework for Dialect Clustering

2.1 Data Representation and Feature Engineering To perform automatic clustering, linguistic data must first be converted into a machine-readable format. This involves the selection of linguistic features, which typically include:

- **Phonetic Features:** The presence or absence of specific phonemes, tonal values, and phonological rules (e.g., the retention of the entering tone).
- **Lexical Features:** Variations in vocabulary for common concepts, often represented using binary vectors or similarity matrices.
- **Grammatical Features:** Syntactic patterns and morphological markers.

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent a set of n dialect sites, where each site x_i is a vector in a d-dimensional feature space. The choice of distance metric

Di, 2000. Historical Linguistics in the Twentieth Century. *Linguistics and Philology*, Issue 10.

Di, 2012. Greenberg' s Genetic Linguistics and the Classification of World Languages: An Overview of *Genetic Linguistics: Essays on Theory and Method*. *Journal of Sino-Tibetan Linguistics*, Issue 6, pp. 37-53. Commercial Press.

Introduction

Joseph H. Greenberg (1915–2001) was a towering figure in 20th-century linguistics, renowned for his pioneering work in linguistic typology and language universals. However, his most significant and controversial contributions lie in the field of genetic linguistics—the study of the historical relationships and genealogical classification of the world' s languages. This article provides an overview and analysis of the core theoretical and methodological frameworks presented in *Genetic Linguistics: Essays on Theory and Method*, a posthumous collection edited by William Croft that synthesizes Greenberg' s lifelong pursuit of a comprehensive classification of human languages.

The Theoretical Foundation of Genetic Linguistics

Greenberg' s approach to genetic linguistics represents a departure from the traditional comparative method that dominated the 19th and early 20th centuries. While the traditional method relies heavily on the reconstruction of proto-languages through rigorous sound laws, Greenberg argued that the primary task of the historical linguist is classification. According to Greenberg, classification must precede reconstruction; one cannot effectively reconstruct a proto-language without first determining which languages belong to the family in question.

Mass Comparison (Multilateral Comparison)

The cornerstone of Greenberg' s methodology is "mass comparison," later referred to as "multilateral comparison." This method involves the simultaneous examination of a large number of languages across a wide geographic area, focusing on basic vocabulary and grammatical morphemes. Greenberg contended that by looking at many languages at once, the "noise" of chance resemblances and borrowings is filtered out, leaving behind the "signal" of genuine genetic relationship.

The logic of multilateral comparison is probabilistic: while two languages might share a similar word for "water" by chance, the probability of dozens of languages across a continent sharing a consistent pattern of similarities in their core lexicons is infinitesimally small, unless they share a common ancestor.

Criteria for Genetic Relationship

Greenberg identified three primary types of evidence used to establish genetic relationships:

1. **Lexical Resemblances:** Similarities in basic vocabulary (e.g., body parts, low numerals, natural phenomena) that are least likely to be borrowed.
2. **Grammatical Evidence:**

On the Genetic Relationship between the Japanese and Korean Languages

Introduction

The question of whether the Japanese and Korean languages share a common origin has long been a subject of scholarly inquiry. In this study, I aim to demonstrate that these two languages are genetically related, belonging to the same linguistic family. By examining their grammatical structures, phonetic correspondences, and core vocabulary, we can discern a profound historical connection that transcends mere geographical proximity.

Grammatical Affinities

One of the most compelling arguments for the genetic relationship between Japanese and Korean lies in their strikingly similar morphological and syntactic frameworks. Both languages are characterized by an agglutinative structure, where grammatical relationships are expressed through the addition of suffixes to invariant roots.

Furthermore, the word order in both languages follows a strict Subject-Object-Verb (SOV) pattern. This consistency extends to the placement of modifiers, which invariably precede the nouns they qualify. The use of postpositional

particles to indicate grammatical cases—such as the nominative, genitive, and accusative—is another shared feature that points toward a common ancestor. For instance, the functions of the Japanese particles *ga*, *no*, and *wo* find direct parallels in the Korean particles *i/ga*, *ui*, and *eul/reul*.

Phonological Correspondences

Beyond structural similarities, a systematic comparison of the phonetic systems of Japanese and Korean reveals regular sound correspondences. While centuries of independent development have led to significant divergence, certain patterns remain discernible.

Through a rigorous analysis of ancient texts and dialectal variations, we can identify cognates that have undergone predictable phonetic shifts. These correspondences are not random but follow specific rules of transformation, which is a hallmark of languages sharing a genetic lineage. For example, certain initial consonants in archaic Korean correspond consistently to specific sounds in Old Japanese, suggesting a shared phonological heritage.

Lexical Comparisons

While loanwords from Chinese have heavily influenced the vocabularies of both nations, the core "native" lexicons of Japanese and Korean provide the most reliable evidence for their relationship. By stripping away the layers of Sinitic influence, we encounter a substratum of basic terms—including pronouns, numerals, and words for natural phenomena—that exhibit clear similarities.

As shown in , the fundamental vocabulary related to the human body and essential actions reveals a degree of similarity that is unlikely to have resulted from chance or borrowing alone. These "swadesh-like" core terms are typically

[Japanese and Korean]. Toyo Kyokai Research, *Academic Report of the Toyo Kyokai Research (Volume 1)*, pp. 159-205.

Ramstedt, G. J. (2004/1957). *Introduction to Altaic Linguistics* (J. Zhou, Trans.). Hohhot: Inner Mongolia Education Press.

How to Distinguish Languages and Dialects: An Exploration of Distance Calculation Methods Based on Core Vocabulary

Abstract

The distinction between "language" and "dialect" is a fundamental yet complex issue in linguistics. This study explores a quantitative approach to this problem by calculating linguistic distances based on core vocabulary. By analyzing lexical similarity and phonetic divergence across a range of linguistic varieties, we propose a systematic framework for establishing objective criteria in language classification. This method aims to supplement traditional qualitative

assessments with empirical data, providing a more rigorous basis for linguistic strategic planning and dialectological research.

---

1. Introduction

The question of how to distinguish between a "language" and a "dialect" has long been a subject of debate in both theoretical and applied linguistics. While the distinction often carries significant political, cultural, and social implications, a purely linguistic definition remains elusive. Traditional criteria, such as mutual intelligibility, are often subjective and difficult to measure consistently. In the context of language strategy and planning, establishing a scientific and reproducible method for classification is of paramount importance.

This paper proposes a quantitative method based on core vocabulary to measure the distance between linguistic varieties. By focusing on the most stable elements of a lexicon—the core vocabulary—we can minimize the interference of recent loanwords and cultural shifts, thereby revealing the underlying genetic or structural relationships between speech varieties.

2. Methodology

2.1 Selection of Core Vocabulary

To ensure the reliability of the distance calculation, we utilize a standardized list of core vocabulary, typically derived from the Swadesh list or its modified versions. These words represent universal human concepts (e.g., body parts, natural phenomena, basic actions) that are least likely to be replaced over time.

2.2 Distance Calculation Formulas

The core of our approach involves calculating the linguistic distance (D) between two varieties. We define the lexical distance based on the proportion of cognates and the degree of phonetic similarity between corresponding forms.

Let L_1 and L_2 be two linguistic varieties. The distance can be expressed as a function of shared features:

$$D = 1 - \frac{C}{N}$$

where C represents the number of cognate pairs and N represents the total number of items compared. To further refine this, we incorporate phonetic distance using Levenshtein distance or similar string-alignment algorithms to account for sound changes.

2.3 Data Processing and

Zhu, 1990, *A Dictionary of Mongolic Languages* [蒙古语族语言词典]. Xining: Qinghai Ethnic Publishing House.

Sun Hongkai, editor-in-chief. 2009. *The Concise Descriptions of the Languages of China' s Ethnic Minorities* (Revised Edition, Vol. 6). Beijing: Ethnic Publishing House.

Sun, Hongkai, Ting Pang-hsin, Jiang Di, and Yan Haixiong. 2017. *Phonology and Vocabulary of the Sino-Tibetan Languages* [汉藏语语音和词汇]. Beijing: Ethnic Publishing House.

The Theories of Altaic Language Family and Key Issues

Wu Hongwei (1996) *Manchu Studies*, Issue 2, pp. 39-54.

Introduction

The study of the Altaic language family has long been a central topic in historical linguistics, particularly concerning the genetic relationships between the Turkic, Mongolic, and Tungusic language groups. This paper aims to review the theoretical foundations of Altaic linguistics and address the critical points of contention that remain at the forefront of the field.

1. Theoretical Foundations of the Altaic Hypothesis

The Altaic hypothesis posits that the Turkic, Mongolic, and Tungusic languages share a common ancestor, referred to as Proto-Altaic. This theory is built upon several pillars of linguistic evidence, including systematic phonological correspondences, shared morphological structures, and a core vocabulary that appears to be cognate across the three branches.

Historically, the development of Altaic studies can be traced back to the early observations of structural similarities, such as agglutinative morphology and vowel harmony. However, the rigorous application of the comparative method in the 20th century provided a more scientific basis for the hypothesis. Scholars have worked extensively to reconstruct the phonological system of Proto-Altaic, focusing on the development of initial consonants and the evolution of the vowel system.

2. Key Issues and Debates

Despite the progress made in Altaic linguistics, several "focal problems" continue to spark debate among researchers. These issues primarily revolve around the distinction between inherited genetic traits and features acquired through long-term linguistic contact.

2.1 Genetic Relationship vs. Areal Diffusion The most significant challenge to the Altaic hypothesis is the "Anti-Altaic" view, which suggests that the similarities between these languages are the result of extensive borrowing and convergence over millennia rather than a common origin. Critics argue that the shared vocabulary is largely restricted to cultural terms and that the morphological parallels are not sufficiently unique to prove a genetic link.

2.2 The Inclusion of Korean and Japanese Another major point of discussion is whether the Altaic family should be expanded to include Korean and Japanese. While some scholars provide evidence for a "Macro-Altaic" group, others remain skeptical, citing the difficulty of establishing regular sound correspondences and the possibility of independent development or different substrate influences.

2.3 Morphological Correspondences Morphology remains a critical battleground for the Altaic theory

Classification of the Kam-Tai Language Family Based on Levenshtein Distance

Zhao Zhijing, Jiang Di (2018). ***Computer Engineering and Applications*****, Issue 19, pp. 62-67.**

Abstract

This study utilizes the Levenshtein distance (LD) algorithm to calculate the linguistic distances between 45 languages and dialects within the Kam-Tai (Dong-Tai) language family. By analyzing a standardized set of core vocabulary, we establish a quantitative framework for assessing the phonetic similarities and divergences among these linguistic varieties. The resulting distance matrix is subjected to computational clustering analysis to generate a phylogenetic classification. Our findings provide a quantitative perspective on the internal subgrouping of the Kam-Tai family, offering a data-driven complement to traditional comparative-historical linguistics.

---

1. Introduction

The Kam-Tai language family (also known as Kra-Dai) represents a significant linguistic group in Southeast Asia and Southern China. Traditionally, the classification of these languages has relied on the comparative method, focusing on shared innovations in phonology, morphology, and lexicon. While effective, traditional methods can be subjective and qualitative. With the advancement of computational linguistics, quantitative methods such as the Levenshtein dis-

tance (LD) have become increasingly prominent for measuring linguistic proximity.

The Levenshtein distance, or edit distance, measures the minimum number of operations (insertions, deletions, or substitutions) required to transform one string into another. In the context of linguistics, it serves as a robust metric for phonetic similarity between cognates. This paper applies this computational approach to 45 representative languages and dialects of the Kam-Tai family to explore their genetic relationships and subgrouping.

2. Methodology and Data

2.1 Data Sources

The primary data for this research consists of a core vocabulary list collected from 45 Kam-Tai linguistic varieties. These include major languages such as Zhuang, Thai, Lao, and Kam (Dong), as well as various local dialects. To ensure consistency, we utilized a standardized set of 200 basic Swadesh-style items, transcribed into a uniform phonetic notation.

2.2 Levenshtein Distance Calculation

The core of our computational analysis is the Levenshtein distance algorithm. For any two phonetic strings S_1 and S_2, the distance $L(S_1, S_2)$ is defined as the minimum cost of transforming S_1 into $

Benedict, Paul K. 1990. Japanese/Austro-Tai. Ann Arbor: Karoma Publishers.

Boller, Anton. 1857. Die Übereinstimmung der Tempus- und Modus-Caraktere in den ural-altaischen Sprachen. SWA 22, 1857, 223-263.

Chisholm, Hugh. Castrén, Matthias Alexander. 1911. Ural-Altaic languages. In Encyclopædia Britannica (11th editon),Vol. 5, 443. Cambridge: Cambridge University Press. (Britannica Editors. "Ural-Altaic languages" . Encyclopedia Britannica, 11 May. 2011, https://www.britannica.com/topic/Ural-Altaiclanguages [Accessed October 6, 2024]).

Clauson, Gerard 1962. Turkish and Mongolian Studies. Royal Asiatic Society of Great Britain and Ireland; sold by Luzac.

Greenberg, Joseph. 2000. Indo-European and Its Closest Relatives: The Eurasiatic Language Family.

Vol. 1: Grammar, Vol.2: Lexicon. Stanford: Stanford University Press.

Holman, Eric W., et al. 2011. Automated dating of the world's language families based on lexical similarity.

Current Anthropology 52, 6: 841-875. Miller, Roy Andrew. 1971. Japanese and the other Altaic languages. Chicago: University of Chicago Press.

Piispanen, Peter Sauli. 2013. Further lexical borrowings from (Pre-)Yakut into the Yukaghiric languages.

Turkic Languages 17, 115–139. Poppe, Nicholas. 1965. Introduction to Altaic linguistics, Harrasowitz, Wiesbaden.

Ramstedt, Gustaf John. 1924. A Comparison of the Altaic Languages with Japanese. Transactions of the Asiatic Society of Japan. 2nd ser.1.41-54.

Robbeets, M. et al. 2021. Triangulation Supports Agricultural Spread of the Transeurasian Languages, Nature, vol.599: 616-621.

Robbeets, Martine, Savelyev, A. 2020(ed.) The Oxford Guide to the Transeurasian languages. Oxford:

Oxford University Press. Robbeets, Martine. 2015. Diachrony of Verb Morphology: Japanese and the Transeurasian Languages.

Berlin: Mouton de Gruyter. Robbeets, Martine. 2005. Is Japanese Related to Korean, Tungusic, Mongolic and Turkic? Wiesbaden:

Otto Harrassowitz Verlag. Sagart, Laurent. 2011. How many independent rice vocabularies in Asia? Rice And Language Across Asia:

Crops, Movement, And Social Change, Sep 2011, Ithaca, United States. .

Sampson, Geoffrey 2022. Review on The Oxford Guide to the Transeurasian Languages. LINGUIST List Sinor, D. 1988. The Uralic Languages: description, history, and foreign influences. New York: Brill Publishers.

Starostin, Sergei. 2003. Co-authored with Anna V. Dybo and Oleg A. Mudrak. Etymological Dictionary of the Altaic Languages, 3 volumes. Leiden: Brill.

Unger, J.M. 1990. Japanese and What Other Altaic Languages? Linguistic Change and Reconstruction Methodology, ed. Philip Baldi, pp. 547-61. New York: Mouton de Gruyter.

Zheng, Tian, et al. 2022. Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative Arising from Robbeets et al. bioRxiv preprint doi: https://doi.org/10.1101/2022.06.09.495471.

Completed September 27, 2025; Revised February 12, 2026

From Altaic to Transeurasian Languages: Traditional Genealogical Classification and Algorithmic Clustering JIANG Di and MENG Wen

Abstract

Studies of the Altaic language family and the recently proposed

Transeurasian classification have traditionally relied on the historicalcomparative method. Despite extensive research, considerable controversy persists,

largely because it remains difficult to determine whether shared linguistic features reflect common inheritance or contact-induced borrowing.

Therefore, a quantitative comparative experimental approach, which uses mathematical techniques and computational models to calculate relational distances among languages, is adopted. The aim is to assess whether the Altaic languages can be considered a language cluster exhibiting phylogenetic affinities.

The analytical framework employs outgroup selection and rooting techniques within clustering experiments to evaluate the Ural-Altaic hypothesis, the traditional Altaic hypothesis, and the Transeurasian hypothesis by means of ingroup-outgroup classification. The results indicate that these related hypotheses are difficult to substantiate. The study compiles a dataset of nearly one hundred languages or dialects from the Indo-European, Uralic, Altaic, Caucasian, Koreanic, Japonic, Sino-Tibetan, and Austronesian language families, and conducts cross-family calculations of similarity-distance relationships among languages and language groups. In addition, Chinese dialects, languages of the Uto-Aztecan family of South America, and languages of the Dravidian family of South Asia are employed as outgroups for infering phylogenetic branching order.

The computational models are based on Levenshtein distance measures and phylogenetic clustering analyses implemented using MEGA. Methodologically, the study introduces criteria related to the number of languages sampled and adopts more radical experimental designs to evaluate the logical consistency of the results. Ultimately, the objective computational analysis refuted the UralAltaic and Transeurasian hypotheses, along with other related proposals, while failing to substantiate the traditional Altaic hypothesis. At the same time, it proposes that a language family with phylogenetic affinity may exist in northeastern Eurasia, namely a Mongolic-Manchu-Tungusic language family (Mongolic-Tungusic).

Keywords

clustering experiment; Levenshtein distance; quantitative comparative experiment; Altaic language family; Transeurasian languages

Note: Figure translations are in progress. See original paper for figures.