

0. Introduction

This book offers a new approach to the long-standing problem of the genealogical affiliation of the Nuristani languages, a small group of closely related languages spoken in the Eastern Hindu Kush around the border of Afghanistan and Pakistan, within the Indo-Iranian subgroup of Indo-European. Since the early 20th century, these languages have been acknowledged as representing a unique Indo-Iranian lineage that does not obviously belong into either of the established subgroups Iranian and Indo-Aryan.

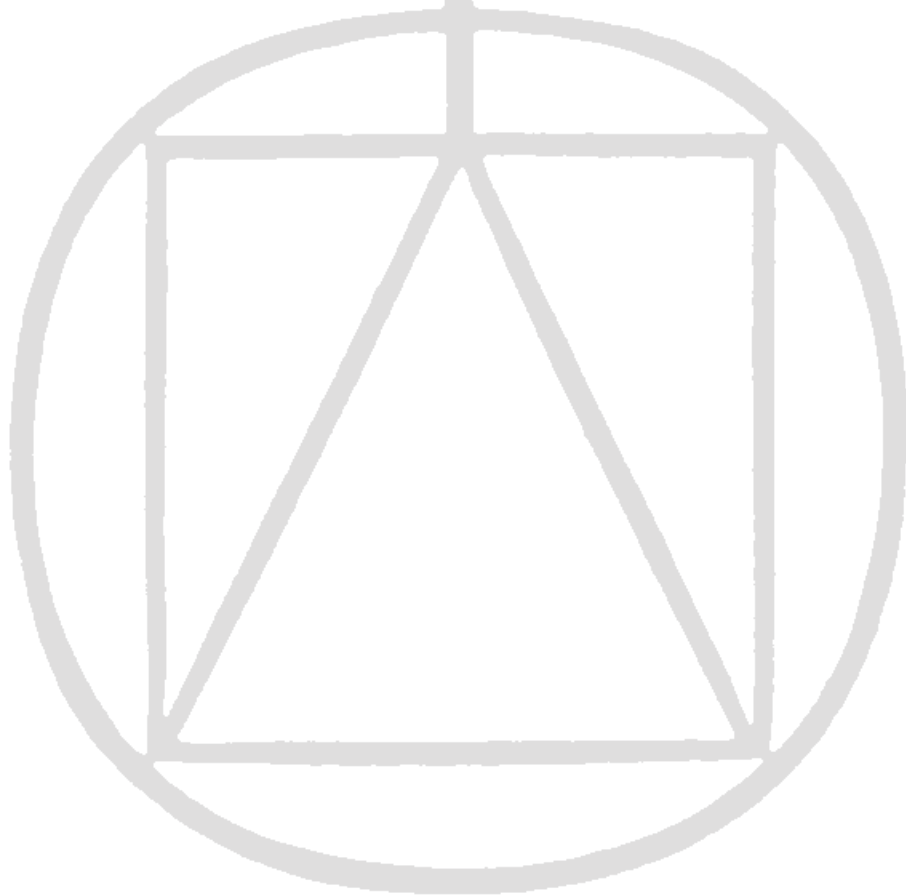
In order to come closer to a reliable classification of the Nuristani group, the topic is approached both via theoretical considerations on language diversification (Section 1) and via a step-by-step examination of the features defining the established groupings Indo-Iranian, Iranian and Indo-Aryan, with a focus on those that are relevant to the question at hand (Sections 2–5). The major part of the work then deals with the features of the Nuristani group itself, examining most of the crucial isoglosses that have been discussed in previous research.

Among Indo-Europeanists, the debate about the classification of Nuristani has remained centered around the same few lexical items presented by Morgenstierne (1926: 50–69; 1945; 1973a) and Buddruss (1977a). Lexical resources published more recently, both by European and North American researchers (e.g., Degener 1998; Strand 1999b; Buddruss & Degener 2015) and by native speakers in Afghanistan and Pakistan (e.g., Tāza 2017; Sun-Aro 2022),¹ have not yet been sufficiently brought to bear on the question. Data from these sources have been fully integrated into the present treatment. Its novelty therefore lies in the fact that it draws on more ample and reliable lexical resources than those which were available to previous researchers.

The conclusions reached in this way differ from some of the more recently published contributions on the topic in suggesting that the Nuristani languages are historically more closely affiliated with the Iranian than the Indo-Aryan subgroup

¹ The dictionaries produced by native speakers in Afghanistan are quite difficult to access in other countries. They are of varying reliability with regard to translation of lemmas and transcription. In some cases, no more than a vowelless ad-hoc orthography in Arabic script is offered. Though they must therefore be approached with some critical thinking, they are nevertheless usually more reliable than the field transcriptions of earlier foreign researchers, who usually had an imperfect command of the languages. I am grateful to Sviatoslav Kaverin for making the dictionaries of Tāza (2017) and Sun-Aro (2022) and Georg Buddruss's unpublished materials on the Ashkun dialect of Wāmā, available to me.

of Indo-Iranian, though they must have been isolated from the Iranian continuum early on and subsequently have come under intense contact influence from Indo-Aryan languages, which led to extensive lexical borrowing, some shared areal sound changes and structural/typological convergence.



1. Preliminaries on Language Diversification

The observation that languages diversify over time, eventually breaking up into families of daughter languages, is one of the fundamental insights that provided the foundation for modern (historical) linguistics. How exactly this diversification proceeds has been a subject of controversy since the beginnings of the field (see, e.g., Schuchardt 1870; Schmidt 1872: 27–28; against Schleicher 1861: 6–7) and the debate especially over how to best represent this process in abstract models (e.g. in the form of trees, waves, chains, networks etc.) has not abated up to the present (see, e.g., Jacques & List 2019; against François 2014; Kalyan & François 2018; and the response in Kalyan & François 2019). Part of the issue is certainly that the process of diversification is most apparent in its results, but difficult to observe while it happens, especially since the time scale involved is larger than the lifespan of individual human beings.

Despite the continued debate over modeling, which is essentially concerned with the appropriateness of this or that metaphor and its respective methodological (dis)advantages,² a general understanding of the ground reality has been available at least since Paul (1886), who understood diversification as emerging from individual idiolectal variation, which is compounded or restricted by the intensity of interpersonal contact:

The life of a language is not conceivable without constant differentiation. If it were imaginable that the languages of individuals in the area of one language were completely alike at some point, the first step to the development of differences among them would surely be made in the next instance. The spontaneous evolution of each one of them must take a particular course based on the particularities of the predisposition and experiences of its bearer. The influence that the individual exerts or suffers never extends further than to a fraction of the collective, and within this fraction there are significant differences in grade. Consequently, a constant leveling of the emerged differences does take place, which consists in divergences from the previous custom being repressed, or, on the other hand, transferred to individuals who had not developed them spontaneously. However, this leveling never becomes complete. It comes close to this only within

² Cf. already Schmidt (1872: 28): “Bilder haben in der wissenschaft nur ser geringen wert” (“metaphors/images have only very little value in science”).

a circle in which constant active contact takes place. The less intense the contact, the more differences can form and be maintained. The possibility of diversification goes even further when there is no longer any contact at all, but only indirect transmission via intermediaries. (Paul 1886: 38; author's translation)³

It is generally assumed that languages usually diversify first into dialects and eventually into separate languages (though the cut-off between the two is obviously arbitrary) and that subgroups eventually emerge from the breakup of separated daughter languages. If this is the case, it implies that the breakup of a single ancestor language is usually followed by a dialect continuum phase, in which at least adjacent varieties still form a community of shared communication. Sections of the continuum may undergo shared innovations and these can either form the basis for divergence into a new language or subgroup, or they might be overlaid by differently distributed innovations that turn out to be more numerous or significant in (preventing) communication in the long run.

Where sharply distinct varieties have not yet come into being, it is unrealistic to consider each innovation that spreads only to one part of the continuum as producing a phylogenetic split. There can be no meaningful distinction between contact spread and internal innovation when divergences have not yet reached the point where mutual understanding between varieties is impeded. In such situations – which are not rare – even “true” shared innovations (as opposed to independent parallel innovations) can arise with incongruent geographic distributions throughout the continuum (*contra* Jacques & List 2019: 140–142; cf. also Kalyan & François 2019: 169–170). If the continuum eventually breaks up into separate languages (and these become subgroups after further diversification), the older isoglosses from the continuum period may be distributed over subgroups in ways that seem unexpected when considered from a later perspective.

³ “Ohne fortwährende differenzierung kann das leben einer sprache gar nicht gedacht werden. Wäre es denkbar, dass auf einem sprachgebiete einmal alle individualsprachen einander vollständig gleich wären, so würde doch im nächsten augenblicke der ansatz zur herausbildung von verschiedenheiten unter ihnen gemacht werden. Die spontane entwicklung einer jeden einzelnen muss nach den esonderheiten in der anlage und den erlebnissen ihres trägers eine besondere richtung einschlagen. Der einfluss, den der einzelne übt oder erleidet, erstreckt sich immer nur auf einen bruchteil der gesamtheit, und innerhalb dieses bruchteils finden bedeutende gradverschiedenheiten statt. Demgemäss findet zwar auch eine immerwährende ausgleichung der eingetretenen differenzierungen statt, die darin besteht, dass abweichungen von dem bisherigen usus entweder zurückgedrängt werden oder aber auf individuen übertragen, die sie spontan nicht entwickelt haben. Diese ausgleichung wird aber nie eine vollständige. Eine annähernde wird sie immer nur innerhalb eines kreises, in dem ein anhaltender regen [sic] verkehr stattfindet. Je weniger intensiv der verkehr ist, um so mehr differenzen können sich bilden und erhalten. Noch weiter geht die möglichkeit zur differenzierung, wenn gar kein directer verkehr mehr besteht, sondern nur eine indirecte verbindung durch mittelglieder.”

This leads us to the possibility pointed out by Garrett (2000) that seemingly characteristic innovations found universally in a certain language family or group may turn out to be results of continued convergence of (a section of) a continuum, rather than reflecting reconstructable features of a common ancestor of the group.⁴ Early attestations and outlier evidence are therefore crucial in accessing the real history and chronology of changes. Garrett (2000: 148–149) discusses the example of the Ancient Greek dialects, some features of which (e.g. the loss of labiovelars) used to be projected back to the Proto-Greek common ancestor before the discovery of Mycenaean revealed that they must have been produced by later convergence of all non-Mycenaean dialects, while Mycenaean itself is a part of the original continuum rather than its ancestor, since it shares innovations with only some of the other dialects.

One might equally adduce cases like the metathesis of liquids in Common Slavic vs. the absence of metathesis in Polabian. Had Polabian gone extinct without leaving any traces, linguists would have had to wrestle with the problem of liquid metathesis occurring in all Slavic languages, but taking different forms in different varieties (e.g. Bulgarian *grad* vs. Russian *górod* ‘town’). One might have reached the conclusion that some form of metathesis was already a feature of the common ancestor, or one might have reconstructed unmetathesized forms to explain the variation. The latter possibility, which is the historically correct one, would have seemed much less plausible without the outlier testimony of Polabian (*gord* ‘town’).

What seems to be necessary for the genesis of a separate language from a continuum of dialects is that innovations accumulate in a group of varieties in such a way that most communication with other related varieties is prevented.⁵ This occurs most easily where natural or social barriers or emigration impede continued contact. This is how a “split” is typically imagined in historical linguistics. Babel et al. (2013: 447), based on Ross (1988) and Pawley & Ross (1995), refer to this process as “network breaking”. It is, however, not the only way how a dialect continuum can break up into separate languages. One might also imagine the case where a different language spreads into the area of the dialect continuum via immigration and/or language shift, thereby severing the

⁴ Phrased in biological jargon, this means that “apomorphic taxa need not be clades” (Babel et al. 2013: 448), i.e., groups with shared features (*taxa*) that are divergent (*apomorphic*) do not have to form a group of descent from an exclusive common ancestor (a *clade*). In other words, shared innovations do not automatically imply the historical existence of an exclusive common ancestor.

⁵ The term “varieties” here can be understood as reaching up to the level of idiolects.

contact relations between related varieties, allowing them to separately accumulate innovations. A variant of this situation occurs when the spreading language is a variety originally situated on one end of the continuum. In this case Babel et al. (2013) use the term “network pruning”, stating that

sharply distinct languages and the eventual family-tree effect can arise out of an earlier dialect network when expanding dialects replace their neighbors. If enough intermediate dialects are pruned, the remaining dialects will be sharply distinct⁶ (Babel et al. 2013: 447).

That is, when innovations have accumulated in a variety on one end of a continuum and speakers of adjacent dialects shift to this variety, two formerly distant sections of a continuum may become adjacent, with the result that communication is prevented and that each of the two goes on accumulating further innovations on its own. If the expanded variety diversifies again after this, it produces a new subgroup.

With these general observations in mind, it is possible to turn to the main topic of this work, the Indo-Iranian family.

⁶ Babel et al. (2013) draw this observation from Schmidt (1872: 28), who presented it in the context of a debate over the internal structure of the Indo-European family. Though Schmidt’s argument aimed at replacing the tree model, particularly in its application to Indo-European, it should be noted that the recognition that “network pruning” is a possible historical process – which seems difficult to deny – does not automatically imply that this is the only way how languages diversify or invalidate the possibility of diversification into subgroups via continuous divergence.

2. The Indo-Iranian Family

Indo-Iranian is the most widely accepted higher-level subgroup of Indo-European and its status as an intermediate node in the diversification of Indo-European is uncontroversial. This is because the Indo-Iranian languages share a number of clear innovations, especially in phonology,⁷ that set them off as a whole against the rest of the Indo-European family. The most important exclusive innovations are (cf. Kümmel 2022: 246–251):

- Brugmann's law: lengthening of PIE **o* in open syllables > **ō* (> **ā* via the Indo-Iranian vowel merger)
 - e.g., OIA *jānu-*; YAv. *zānu-*; NKal. *zā* < PIE **ǵonu-* 'knee' (> Greek γόνυ)
- Vowel merger: PIE **ā, *ē, *ō* > PIr. **ā*
 - e.g., OIA *catvārah, bhrātar-*; YAv. *čaθvārō, brātar-*; NKal. *čatā, brā* < PIE **k^wetwores* 'four', **b^hreh₂ter-* 'brother' (> Greek τέσσαρες, φράτηρ)
- Epenthetic vowel **i* next to laryngeals (preserved in somewhat distinct distributions in the daughter languages)

There are also some innovations that are clearly present in Indo-Iranian, but not exclusive to it, making it debatable whether their occurrence in other Indo-European subgroups is shared or merely parallel with Indo-Iranian, like the following:

- Satemization (Ćatamization): PIE palatovelars **k̑, *g̑^(h)* > affricates **č, *ǰ^(h)*
 - e.g., OIA *dása-*; YAv. *dasa-*; Kt. *duč* < PIE **deǵm* 'ten'
- Palatalization of velars: PIE labiovelars & plain velars **k^(w), *g^(w)* > affricates **č, *ǰ* before PIE **ē* and **ī* (this chronologically precedes the Ir. vowel merger)
 - e.g., OIA *pāñca-*; YAv. *panča-*; NKal. *pūč* < PIE **penk^we-* 'five'
- Syllabic nasals PIE **m̥, *n̥* > PIr. **a*
 - e.g., OIA *śatām*; YAv. *satəm* < PIE **k̑mtom-* '100'

For some other isoglosses, the classification as Indo-Iranian innovations is possible, but not beyond doubt:

- Bartholomae's law: Progressive assimilation in obstruent clusters beginning with a voiced aspirate, e.g., PIE **b^hud^h-to-* > OIA *buddha-* 'awoken'

⁷ Kümmel (2022: 250) mentions only one potential exclusive morphological innovation: the use of the "weak" stem of nouns in the accusative plural.

- The rule is still applied in Old Avestan, despite the deaspiration of voiced aspirates, but is lost in later Iranian; it is unclear whether it is an Indo-Iranian innovation or an archaic feature that has been lost in other subgroups of Indo-European (Kümmel 2022: 246–247)
- Grassmann’s law: Dissimilation of voiced aspirates when preceding another voiced aspirate
 - Visible only in Indo-Aryan, since voiced aspirates are deaspirated in Iranian and Nuristani (Kümmel 2022: 247–248)
- Liquid merger: PIE $*l + *r > *r$
 - Debated due to the existence of a phoneme /l/ in (especially later) OIA and later Iranian, but the original PIE distribution of $*l$ and $*r$ is in any case lost everywhere, making an original merger likely, though several open questions remain (see Mayrhofer 2002)
- RUKI rule: PIE $*s > \text{PIIr. } *š$ after PIE $*r/*l, *u, *k$ and $*i$
 - Application in Nuristani is debated, see Section 6.3.
- Aspiration from the influence of laryngeals
 - The results are phonetically aspirated only in Indo-Aryan, but fricatives in Iranian, and Nuristani has no traces of aspiration. Karim (2021: 9–10) and Kümmel (2022: 250–251) therefore suggest that the Iranian instances could also be explained separately as results of the general Iranian pre-consonantal fricativization. However, fricativization before obstruents and before laryngeals is perhaps not very likely to result from the same phonetic process and there is at least one development that suggests presence of voiceless aspirates in the prehistory of Iranian ($*TVND^h > *T^hVND$; see the discussion in Section 5).

In addition to these classical examples of phonological innovations, there are also some more idiosyncratic divergences that make a single intermediate ancestor language plausible as a historical reality.⁸ A number of common Indo-European words appear in slightly unexpected forms:

- ‘tongue’: OIA *jihvā́-*, YAv. *hizuuā-*, Kt. *dič* vs. Gothic *tuggō* which is more expected from PIE $*d̥ḡʰueh_2s$. Though most descendants of this word have undergone

⁸ François (2014: 178) stresses the potential value of what he calls “lexically-specific sound changes” for linguistic classification, since – as he claims – such sporadic divergences are more likely to be transmitted as such only within a single community. Jacques & List (2019: 144–146, 148) do not dispute this observation in principle, but they are right to caution that the probative value of such divergences depends on how each particular case actually came about. Thus, e.g., idiosyncratic divergences resulting from variously levelled proto-variation are not strong evidence for shared descent.

- irregular deformations, the Indo-Iranian languages agree in reflecting an unexpected *i* in the first syllable. The onsets appear to be variously dissimilated;
- ‘tear’: OIA *ásru-*, YAv. *asrū-*⁹, Kt. *ač* with vowel onset vs. Greek δάκρυ, English *tear*, but the initial plosive is missing also in Agnean (Tocharian A) *ākār* and Lithuanian *āšara*;
 - ‘heart’: OIA *hřd-*, *hřdaya-*, OAv. *zərəd-*, YAv. *zərədaia-*, Kt. *zěré* pointing to an onset with virtual PIE **ǵh-* vs. all other Indo-European languages pointing to PIE **k-* (e.g. Latin *cor*, Armenian *sirt*). Since the expected onset *ś* does appear in OIA *śraddhā-* ‘faith, trust’ and *śrād dhā-* ‘to trust’ < PIE **kred d^heh₁-* lit. ‘to set the heart’ (but OAv. *zrazdā-* ‘trusting’ etc.), it is likely that the explanation should be sought in a synchronic alternation of the Indo-Iranian ancestor language, which underlines the plausibility of its existence as a real language system.⁹

There are also a number of words which seem to have been inherited from the common ancestor of Indo-Iranian, but find no correlates in other Indo-European languages, which makes it likely that they entered into the Indo-Iranian proto-language as loanwords. Following Lubotsky (2001a), some have sought to associate these words with an unattested language supposedly spoken by the bearers of the Bactria-Margiana Archeological Complex in Central Asia. Another notable lexical agreement is the reconstructable self-designation **āřia-* shared among the earliest Indo-Iranian languages (OIA *ārya-*, YAv. *airiia-*, OP *ariya-*).

These features all speak in favor of the historical reality of a single Proto-Indo-Iranian language, from which all later Indo-Iranian languages descend. More difficulties in subclassification are encountered below the umbrella of Indo-Iranian. While the scientific consensus for a long time had been that there are two clearly distinguishable subgroups named Iranian and Indo-Aryan, and that Iranian itself must be divided into two (East and West Iranian) and further into four subgroups (Northeast, Southeast, Northwest and Southwest Iranian), more recent research has raised doubts about almost all aspects of this model. At least the mentioned scheme of subclassification for Iranian is by now widely accepted as untenable (cf. Korn 2016; 2019) and no alternative scheme has since gained universal acceptance. Instead, even the idea that Iranian as a whole forms a coherent subgroup descending from an intermediate common ancestor has been called into

⁹ The alternation in question might have been a PIr. sandhi pattern paralleling the OIA rules *-d, -t + ś- > -c# ch-* and *-d, -t + j- > -j# j-* in the voiced sphere, i.e. **-d + *ć- > *-f# f^h-* next to **-t + *ć- > *-č# č^h-* the conditioning for which would have been lost with the loss of the final voicing contrast in plosives. As a neuter, the PIr. word for ‘heart’ would often have appeared after, e.g., the neuter demonstrative **tad*.

question (Tremblay 2005). While the status of Indo-Aryan as a subgroup is usually not doubted, puzzling questions about its history of divergence from Iranian have emerged out of the fragmentary testimony of the Indo-Iranian adstrate vocabulary of Mitanni (~ 15th c. BCE) and the Nuristani outlier group (reliably attested only since the 20th century CE), which is often tentatively considered a kind of “third branch” of Indo-Iranian.

In the following sections, I will discuss the status of each of the four potential subgroups of Indo-Iranian, starting with Indo-Aryan as the most unproblematic and moving on to Mitanni-Aryan, Iranian and finally Nuristani.

