Henrik Liljegren*

The Hindu Kush–Karakorum and linguistic areality

https://doi.org/10.1515/jsall-2021-2027
Published online April 22, 2021

Abstract: The high-altitude Hindu Kush–Karakoram region is home to more than 50 language communities, belonging to six phylogenies. The significance of this region as a linguistic area has been discussed in the past, but the tendency has been to focus on individual features and phenomena, and more seldom have there been attempts at applying a higher degree of feature aggregation with tight sampling. In the present study, comparable first-hand data from as many as 59 Hindu Kush–Karakoram language varieties, was collected and analyzed. The data allowed for setting up a basic word list as well as for classifying each variety according to 80 binary structural features (phonology, lexico-semantics, grammatical categories, clause structure and word order properties). While a comparison of the basic lexicon across the varieties lines up very closely with the established phylogenetic classification, structural similarity clustering gives results clearly related to geographical proximity within the region and often cuts across phylogenetic boundaries. The strongest evidence of areality tied to the region itself (vis-à-vis South Asia in general on the one hand and Central/West Asia on the other) relates to phonology and lexical structure, whereas morphosyntactic properties mostly place the region's languages within a larger areal or macro-areal distribution. The overall structural analysis also lends itself to recognizing six distinct micro-areas within the region, lining up with geo-cultural regions identified in previous ethno-historical studies. The present study interprets the domain-specific distributions as layers of areality that are each linked to a distinct historical period, and that taken together paint a picture of a region developing from high phylogenetic diversity, through massive Indo-Aryan penetration and language shifts, to today's dramatically shrinking diversity and structural stream-lining propelled by the dominance of a few lingua francas.

Keywords: areal typology; convergence; feature aggregation; language contact; micro-area

***Corresponding author: Henrik Liljegren**, Stockholm University, Stockholm, Sweden,
E-mail: henrik@ling.su.se

Open Access. © 2021 Henrik Liljegren, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

More than 50 distinct ethnolinguistic communities inhabit the mountainous northwestern fringe of the Indian subcontinent. This region, here referred to as the Hindu Kush–Karakorum, is spread over the territories of several countries – primarily Afghanistan, Pakistan and India – and comprises languages belonging to six linguistic phyla: Indo-Aryan (in the majority), Nuristani, Iranian, Sino-Tibetan, Turkic and the isolate Burushaski.[1] The linguistic profile of this region has been the topic of discussion for almost half a century: Is it a linguistic area? Or would it be better described as a transitional zone between areas? Or is it, perhaps, an accretion zone, displaying a high degree of linguistic diversity rather than significant language convergence? The tendency in the past has been to focus on individual features and phenomena, sometimes based on relatively sparse data, and more seldom have there been attempts at applying a more fine-grained analysis. In the present study, these issues are revisited by comparing a tight sample of 59 Hindu Kush–Karakorum language varieties for 80 structural features, based on comparable primary data. The relative frequency of feature values as well as their geographical distributions are investigated.

In Section 2, the geographical region and its linguistic landscape is described in some detail, and the issue of areality, especially as it pertains to the Hindu Kush–Karakorum region, is reviewed and problematized. Section 3 details the data set and the context of data collection, and describes the annotation and coding process and the main analytical choices that were made. Section 4 briefly touches on lexical similarity, based on a subset of the data, and discusses to what extent a comparison of basic vocabulary, particularly when we count shared cognates, lines up with established language classification. In Section 5, which is the most central part of this paper, the results of the present study are presented and discussed, with evaluation of: a) to what extent structural similarity in general lines up with phylogeny, and to what extent it cuts across phylogenetic boundaries; b) whether the similarity measures differ in any significant way between the five language domains considered (phonology, grammatical categories, clause structure, word order and lexico-semantic structure); and c) to what extent we can find a correlation between feature distribution and geography, and thus evidence of linguistic areality in the region. Finally, in Section 6, the findings of the previous section are summarized and put into a larger historical and typological perspective

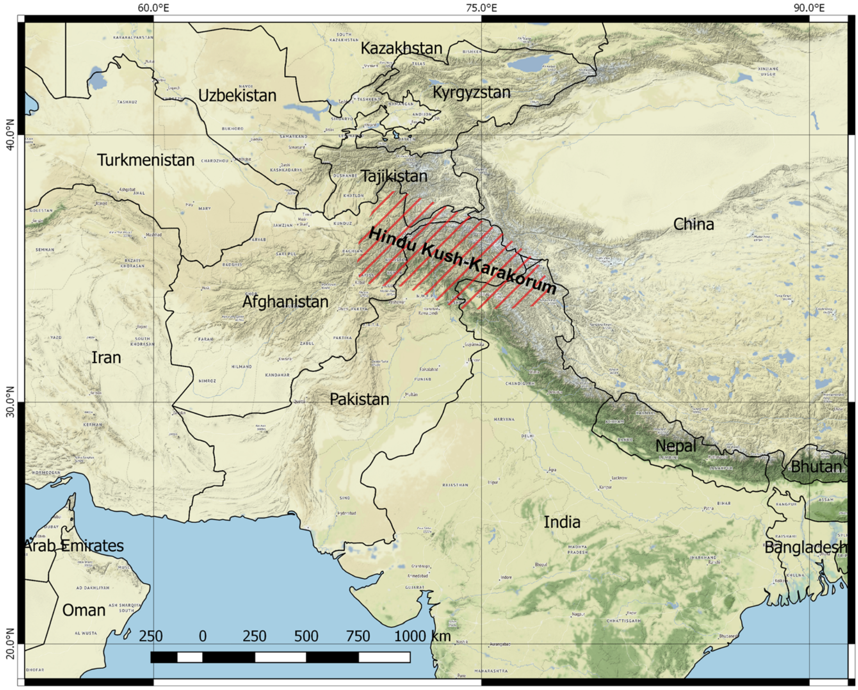
[1] While Indo-Aryan, Nuristani and Iranian are indeed subgroupings within Indo-European, these will for practical purposes be treated here as distinct linguistic phyla, vis-à-vis the the other three – scantily represented – language families.

and some conclusions are reached regarding the areal nature of the Hindu Kush–Karakorum, in the present, as well as in the past, as far as we are able to determine.

2 Background

The Hindu Kush–Karakorum (henceforth abbreviated as HK) is not an exact and established term.[2] It is used here as an approximate cover term for a linguistically diverse region which geographically and culturally lies at the crossroads of South and Central Asia (see map in Figure 1). As a physical region, it is characterized primarily by high altitudes, as compared to the Central Asian steppes, in the north, and the Indo-Gangetic Plains, in the south, and it can also be described as constituting the westernmost end of the larger Himalayan arc. Politically, the HK region has for centuries been divided between several contending powers, and even today a number of international borders traverse this region; it comprises present-day northeastern Afghanistan, the mountainous north of Pakistan, and the northern parts of the two Indian union territories Jammu & Kashmir and Ladakh. The situation is further complicated by the prolonged, and from time to time violent, dispute between India and Pakistan over Kashmir and surrounding areas. However, in spite of the divided status of this region throughout its modern and pre-modern history, historical research asserts that the region once formed a cultural unity (Cacopardo and Cacopardo 2001; Jettmar 1975). Cacopardo and Cacopardo (2001: 25) state that "under a great variety of local forms, a nucleus of common cultural traits characterized the whole region in the past, in the socio-economic as well as in the symbolic domain", in direct reference to pre-Islamic *Peristan*, a geo-cultural entity that largely overlaps with my HK region. Today's HK region has, in contrast, become an integral part of the larger Islamic cultural and religious context of South and Central Asia, although in various forms – Sunni, Shiite, Ismailite and Nurbakhshi. There are only two exceptions to the present Muslim hegemony: The Buddhist, or partly-Buddhist, population of Ladakh, at the eastern fringe of the region; and the tiny non-Muslim section of the Kalasha community residing in a few remote valleys in northern Pakistan's Chitral district (Heegård Petersen 2015: 13), otherwise completely surrounded by groups that were Islamized at the latest between one and two centuries ago.

[2] The difficulty of finding a suitable and practical term to refer to this region in its entirety is illustrated by the use of extreme multi-compounding by other scholars, such as Tikkanen's *Hindu Kush–Kohistan–Karakoram-Pamir* (2008: 258) and Bashir's *Pamir–Hindukush–Karakoram–Kohistan–Kashmir region* (2016: 264).



Map created with QGIS 2.18.26 by Henrik Liljegren 2020. Stamen Terrain map: © Stamen Design, under a Creative Commons Attribution (CC BY 3.0)

Figure 1: The Hindu Kush–Karakorum region in Inner Asia.

Linguistically, the region is multilingual and diverse, and it is the point of geographical convergence of four wide-spread phylogenetic groupings: Indo-Aryan, Iranian, Sino-Tibetan and Turkic. In addition, this region is home to the Nuristani languages and the language isolate Burushaski. The numerically dominant phylogenetic component of HK is Indo-Aryan, comprising about 30 distinct speaker communities, represented in large parts of the region, although in a higher number in the south than in the north. Both the northernmost and the westernmost distribution of Indo-Aryan in South Asia is found within this particular region. Most of those Indo-Aryan languages were previously referred to as "Dardic", sometimes even lumped together with the Nuristani (earlier referred to as "Kafiri") languages as transitional between Indo-Aryan and Iranian. While it has been demonstrated that at least some of those Indo-Aryan languages display significant archaisms (Morgenstierne 1974: 3), on the one hand, and shared contact-related developments (Bashir 2003: 821–822), on the other, there is no justification for collectively grouping these language vis-à-vis other Indo-Aryan languages of the Northwestern zone (Morgenstierne 1961: 139).

As for Iranian, its easternmost reach is found within the HK region and its immediate surroundings. Those languages are primarily spoken in the western half of the region. Iranian is the second largest phylogenetic component of the region, including about 10 languages, but classification-wise there is a slightly higher level of diversity within the Iranian languages represented than there is within the corresponding Indo-Aryan element. Most of them are Eastern Iranian languages for which the label "Pamiri" has been applied, but similarly to the aforementioned term Dardic, Pamiri is as much an areal designation as a truly phylogenetic one (Bashir 2009: 857; Èdel'man and Dodykhudoeva 2009: 776; Wendtland 2009: 173). These latter speaker communities are almost without exception comparatively small. We find them in and around the Pamir Mountains and the Wakhan corridor, also well-represented in neighbouring Tajikistan, and more marginally in adjacent areas of Xinjiang in China. Local varieties of the two major languages, Dari (or Eastern Persian), which is Western Iranian, and Pashto, which is Eastern, are also well-represented within the HK region.

Nuristani, with its 5–6 languages, is spoken in some of the region's most remote valleys in high-altitude northeastern Afghanistan. These languages, all of them well within the boundaries of HK, are in most modern-day treatments considered a separate phylogenetic grouping within the larger Indo-Iranian branch of Indo-European, alongside Indo-Aryan and Iranian. It should be born in mind, however, that there is no consensus among experts on its exact placement in regard to those other two branches (Degener 2002: 103–104). Untangling long-standing areal effects from inheritance is not a trivial matter when comparing Nuristani with e.g. its HK Indo-Aryan neighbours (Degener 2002: 115–116; Zoller 2005: 13–15).

Sino-Tibetan (in this particular case often referred to as Tibeto-Burman, by some seen as a major subgroup of the former) is represented by 3–4 languages at the eastern fringe of HK, in Pakistani-administered Baltistan and in Indian-administered Ladakh, while at the same time constituting a major component of the Himalayas continuing east of the HK region. The Sino-Tibetan varieties of HK are considered part of the Bodish branch, and more specifically Western Tibetan, and in comparison with other Tibetan languages these display significant archaisms as well as shared innovations, the latter obviously as a result of Indo-Aryan contacts (Zeisler 2005: 53–59). Two separate branches of Turkic are represented in the region, both essentially varieties of major Turkic languages of Central Asia, and within HK spoken almost exclusively on Afghan territory. Kyrgyz is northwestern Turkic or Kipchak, and is at home in the remote eastern corner of the Wakhan corridor, with related groups across the border of China. Uzbek is part of the southwestern branch, or Uyghur-Karluk; a southern, Persian-influenced (Reichl 1983: 481) variety of it is spoken in the western parts of Badakhshan. Burushaski, finally, is nowadays a speaker community of a rather modest size and geographical distribution but without doubt represents one of the oldest, if not the most ancient, linguistic layer still vital in the HK region (Tikkanen 1988). It is primarily spoken in a few mountain valleys in the extreme north of Pakistan's Gilgit-Baltistan province. While there is no lack of more or less serious attempts at re-classifying it as belonging to one or the other known linguistic phylum, it is still mostly treated as a language isolate.

Is the HK region a linguistic area? Yes and no. It all depends on how we choose to define the terms "area" and "areal" in this context. If by that we mean a geographical area with well-defined and neat boundaries within which all or most of the languages, regardless of phylogenetic identity, share a significant number of unique features that have arisen as the result of contact, the answer will most certainly be no. This is the definition and traditional usage of the term linguistic area or Sprachbund that is associated with e.g. the Balkan Sprachbund and similar cases.

If we instead mean a convergence zone with a core that share certain linguistic features as the result of many local contact situations that have existed for a prolonged time period, with surrounding subareas in which languages share some, but not all, of the same features, and to a varying extent display other micro-areal convergence patterns, the answer may be yes. The latter reflects a viewpoint akin to that expressed by Dahl concerning areality: "At the most basic level, linguistic contact relationships are binary: one language influences another. An area is then simply the sum of many such binary relationships. But such an area need not display shared features […] that characterize all the languages within the area and which define a clear boundary to the rest of the world" (2001: 1,458). In a similar vein, the results of a comprehensive study of language contacts in the Circum-Baltic area are put forward in the following way: "In the CB area, convergence works primarily on a micro-level. It reflects language contacts of groups of people and maximally, of two or three languages. Convergence that comprises more than two or three languages […] is always the result of the overlapping and superposition of different language contacts" (Koptjevskaja-Tamm and Wälchli 2001: 728). Similar views are expressed in reference to various other recent cross-linguistic studies of particular geographical regions around the world, essentially reflecting a shift in emphasis from counting isoglosses and arguing for well-circumscribed linguistic areas as if they were concrete entities in their own right, to studying linguistic properties and their distributional patterns in a more open-ended way (Muysken 2008: 2; Koptjevskaja-Tamm 2010: 577). The present study aligns itself primarily with this latter perspective.

Contrary to some other prospective convergence areas around the world, such as Mainland Southeast Asia (Enfield and Comrie 2015: 16–18), the HK region has not to this day been subject to a systematic and detailed study, such as the one proposed by Koptjevskaja-Tamm (2010: 584), that applies a high degree of feature aggregation and tight sampling on comparable data to produce a microtypology that measures similarities as well as differences across the region, and in which that microtypology is subsequently evaluated against a broader, global typological background. A number of serious suggestions have indeed been put forward and a few significant pilot studies have been carried out by scholars with good insights into the region's linguistic composition and its typological profile, from the 1970s onward. A "Central Asiatic linguistic area" is proposed by Toporov (1970), based on striking similarities in the phoneme inventories of languages in the region, irrespective of language relatedness. The central role played by Burushaski in this areal distribution, which is emphasized in that work, is echoed by Èdel'man (1980, 1983: 16), who apart from phonological similarities (such as retroflex fricatives and affricates) includes several diagnostic features of this Central Asiatic linguistic area belonging to other language domains, among them the overall presence of a vigesimal numeral system.

Some of the more substantial and well-informed suggestions put forward regarding areality, whether macro-areal with special reference to the HK region, truly areal, or micro-areal, restricted to a sub-region of HK, are those of Bashir (1996a, 1996b, 2003: 823, 2016). Phenomena discussed by her, apart from those already mentioned, include e.g. the prevalence of left-branching syntactic structures, grammaticalized inferentiality, and multi-valued deictic systems. Tikkanen (2008) revisits some of the aforementioned suggestions, particularly those related to shared phonological features, and concludes that the area corresponding to HK constitutes a combination of convergence between two macro-areas (a Central Asian and a South Asian) that happen to overlap precisely here, and a cluster of very ancient micro-areal features integral to the region itself. Concerning the latter, he presents arguments (2008: 258–259), in addition to those discussed elsewhere (Tikkanen 1988, 1999), for substratal origins, both including Burushaski-type languages and the former presence of languages belonging to phylogenies without any present-day representation in HK. A convergence domain in HK that has received comparatively little attention is that of semantics and lexical organization (Koptjevskaja-Tamm and Liljegren 2017: 215–223). Apart from area-wide similarities in segmental phonology, Baart (2014) presents a study that shows tonality or the presence of a pitch-accent system as typical of the languages of the region.

Significant micro-areas or local convergence zones within HK have also been suggested or described. A number of independent linguistic features characterizing languages in the western part of the HK region have been postulated, possibly centered and radiating out of Nuristan, but not confined to Nuristani languages. Di Carlo (2011) discusses the co-occurrence of pronominal kinship suffixes and retroflex vowels, the latter echoing an idea put forward by Mørch (1997). Liljegren and Svärd (2017: 468–470) add the occurrence of bisyndetic contrast marking, and Heegård and Liljegren (2018: 155) add geomorphic coding of spatial reference, both phenomena largely overlapping geographically with the subregion of HK that until only a couple of centuries ago still constituted a pre-Muslim island inside a larger region which had embraced Islam many centuries earlier.

Another perspective, which we will have reason to return to, is that of areal profiling, not so much regarding area-specific convergence as that of characterizing an area in relation to other areas in terms of its genetic density. In Nichols' global differentiation of spread zones and accretion zones, respectively, the Himalayas (or the Pamir-Himalayas) is classified as an accretion zone (1992: 21), described as "an area where genetic and structural diversity of languages is high and increases over time through immigration" (1997: 369).

3 Sampling and data collection

Rather than trying to identify a smaller and representative sample or depending on data from multiple and heterogeneous sources, the aim of this study was to gather primary data from as many as possible of the distinct language varieties spoken in the HK region. This was carried out in a series of multi-language collaborative workshops held in the period 2015–2018 in Islamabad (Pakistan), Kabul (Afghanistan), Faizabad (Afghanistan) and Srinagar (India), supplemented with a few individual, single-language, sessions in Islamabad, Gilgit (Pakistan) and Kargil (India). In total, 79 native language consultants (1–3 for each variety) were recruited and contributed data. Recruitments, local administration and support, translations of elicitation materials and textual data (between English and Dari, Pashto and Urdu, respectively) and digitization were handled in close collaboration with three institutions, one in each country. Comparable data from as many as 59 Hindu Kush–Karakorum language varieties was thus elicited, processed and (partly) analyzed (Table 1 lists the sample varieties; see Appendix for more precise information on locations).

All six previously identified phylogenetic groupings are represented, with a natural predominance of Indo-Aryan. In several cases, more than one variety of what is essentially one language were included. Partly, that was in order to represent national varieties of a language spoken across international borders, such as Pashto of Afghanistan, Pakistan as well as Indian Kashmir, and Wakhi spoken in Afghanistan and in Pakistan. Partly, that was in order to obtain higher resolution regarding some known geographical continua of a language or a subgroup. The Pashai continuum is a case in point, for which as many as nine varieties

Table 1: Language varieties represented in the data sample (listed phylogenetically); language name, specification of speaker location (Afg = Afghanistan; Ind = India; Pak = Pakistan) when more than one variety is represented, and its three-letter ISO code (the author's added differential coding is given within parentheses).

Burushaski (2)

Burushaski, Hunza [bsk (h)]
Burushaski, Nagar [bsk (n)]

Indo-Aryan (33)

Bateri [btv]
Brokskat [bkk]
Dameli [dml]
Gawarbati, Afg [gwt (a)]
Gawarbati, Pak [gwt (p)]
Gawri [gwc]
Gojri, Afg [gju (a)]
Gojri, Pak [gju (p)]
Hindko, north [hno]
Indus Kohistani [mvy]
Kalasha [kls]
Kalkoti [xka]
Kashmiri, Ind [kas (i)]
Kashmiri, Pak [kas (p)]
Khowar [khw]
Kohistani Shina [plk]
Kundal Shahi [shd]
Pahari-Pothwari [phr]
Palula [phl]
Pashai, Alasai [psh (ai)]
Pashai, Alingar [psi (ar)]
Pashai, Alishang [glh (ag)]
Pashai, Amla [psi (am)]
Pashai, Aret [aee (at)]
Pashai, Chalas [aee (ch)]
Pashai, Korangal [aee (kg)]
Pashai, Sanjan [glh (sn)]
Pashai, Shemal [aee (sh)]
Sawi [sdg]
Shina, Gilgit [scl (p)]
Shina, Gurez [scl (i)]
Torwali [trw]
Ushojo [ush]

Iranian (13)

Dari, Darwoz [prs (d)]
Ishkashimi [isk]
Munji [mnj]
Parachi [prc]
Pashto, north, Afg [pbu (a)]
Pashto, north, Ind [pbu (i)]
Pashto, north, Pak [pbu (p)]
Rushani [sgh (r)]
Sanglechi [sgy]
Shughni [sgh (a)]
Wakhi, Afg [wbl (a)]
Wakhi, Pak [wbl (p)]
Yidgha [ydg]

Nuristani (6)

Ashkun [ask]
Kamviri [xvi]
Kati, east [bsh (e)]
Kati, west [bsh (w)]
Prasun [prn]
Waigali [wbk]

Sino-Tibetan (3)

Balti [bft]
Ladakhi [lbj]
Purik [prx]

Turkic (2)

Kyrgyz [kir]
Uzbek, south [uzs]

were included, bearing in mind Morgenstierne's well-informed and yet somewhat reluctant decision to divide the many and challenging varieties into four major groups, or *de facto* "languages", that he named NW, NE, SW and SE Pashai (1967: 12–13). In fact, there are only a few, almost exclusively small, language communities that are not represented at all in this sample, in some cases due to difficulties in recruiting consultants (e.g. Domaaki [dmk], Chilisso [clh] and Tregami [trm]), but in other cases simply because these speech forms most likely have ceased to exist as actively spoken languages (e.g. Shumashti [sts], Tirahi [tra], Grangali [nli], Badeshi [bdz] and Gowro [gwf]). In Figure 2, the 59 sample varieties are plotted on a map of the region.

With minor exceptions, the dataset for each language consists of seven components: three word lists, a sentence questionnaire, a translated parallel text, context-elicited demonstrative expressions, and a stimulus-based narrative. See Table 2 for details on each component. For the large majority of language varieties (53 of 59), data collection was carried out in the context of a 4–5 day workshop, representing five or more languages, in which the following procedure was followed: the participants were given a basic introduction to one of the components, e.g. kinship terms; (if applicable) they were given time to prepare themselves for recording (either individually or group-wise) by filling in a word list or questionnaire in whatever style they preferred: they were then invited, each in turn, to a (makeshift) recording studio to be audio or audio-and-video recorded; and after that a considerable amount of time was spent in comparing and discussing

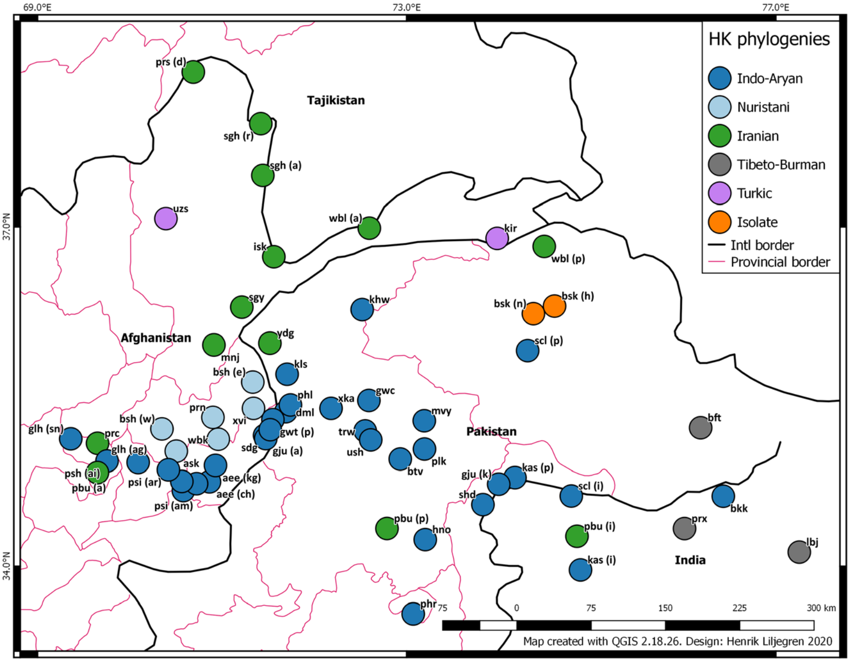


Figure 2: The 59 sample varieties plotted in the HK region (Note that the incomplete boundary demarcation between Pakistan and India is the so-called Line of Control, and does not as such constitute a legally recognized international border).

Table 2: Dataset: components (abbreviations used in data references within square brackets), descriptions and recorded forms.

Data component	Description	Recorded form
40-list [40list]	Word list including the 40 basic vocabulary items used by ASJP (Wichmann et al. 2016)	Written (mostly Arabic-based) + audio
Kinship [Kin]	Word list including 95 kin relations, designed by the author	Written + audio
Numerals [Num]	Word list including the cardinal numerals 1–50, 60, 70, 80, 90, 110, 120, 200, 1,000	Written + audio
Valency [Val]	Sentence questionnaire representing 87 verb meanings, designed by the Leipzig Valency Classes Project (http://valpal.info/about/project)	Written + audio
North Wind [NW]	Translation of the traditional fable *The North Wind and the Sun*, widely used for illustrating the phonetics of numerous languages (International Phonetic Association 1999)	Written + audio (for a subset: written + audio + video)
Demonstratives [Dem]	Expressions used in reference to objects situated at various distances from the speaker. An elicitation kit for this was designed by the author, largely inspired and guided by Wilkins (1999)	Audio + video (for a subset, the consultants transcribed their own utterances based on the audio recordings)
Pear Story [PS]	Natural or semi-natural speech used in retelling the contents of the 6 min "Pear film"; see Chafe (1980) and http://pearstories.org/	Audio + video (for most of the varieties, the consultants transcribed their own speech and provided a translation to a lingua franca)

particular pieces of data among the participants. All consultant-produced written material was either saved electronically or photocopied, to aid further processing.

Subsequent to data collection, the material was organized, transcribed and coded. The data set allowed for extracting a basic word list of 100 comparable meanings as well as for classifying each variety according to 80 binary structural features. Of these, 16 reflect the domain of phonology, another 16 reflect grammatical categories, 16 clause structure features, 16 word order features and 16 features related to lexico-semantic structure. A description of each of the 80 binary features appears in the Appendix. Two control varieties were added when carrying out the structural analysis, namely Urdu and Standard Dari, in order to determine modern-day superstratal effects, as these two languages represent the two most important languages of national scope as far as cross-community communication and media are concerned, Dari in Afghanistan and Urdu in Pakistan and Indian-administered Kashmir.

The illustrative examples that are given in this article are to a large extent drawn from the data set described above. Those examples can be recognized by their references beginning with the language code (as outlined in Table 1), followed by the abbreviation for the data component (in Table 2), then an abbreviation tied to a particular language consultant, and finally an individual item number.[3] In some cases, however, secondary sources have been used, as well as field data collected by the author outside the scope of the present project.

4 Lexical analysis

In a combined attempt at corroborating – or problematizing – previous classifications and checking the data for accuracy, we measured lexical proximity based on the 100-item word list that was extracted (see Appendix for a full list). The items chosen mainly represent body parts, close kinship relations, personal pronouns, lower numerals, basic actions, common substances and a few geographical features and animals. The underlying assumption is that those vocabulary items are among those less likely to be replaced and therefore are inherited to a larger extent than other, less basic vocabulary. By using the language comparison tool Cog (version 1.3.4.10016)[4] to automatically identify cognates, detect sound correspondences and measure lexical similarity, a distance matrix was produced. The higher the number is of shared cognates between any two language varieties, the shorter the distance is between them. This distance matrix in its turn generated the lexical neighbor-joining tree that can be seen in Figure 3. The very general conclusion we can draw from the language clustering that the figure visualizes is that the data largely confirms, or at least does not contradict, established classification into the six major phylogenetic groupings, as referred to in Section 2.

The distinctiveness of the Turkic, the Sino-Tibetan and the Burushaski varieties, respectively, goes without saying. The Nuristani as well as the Iranian

[3] The intention is to gradually make the data set available on the Cross-Linguistic Linked Data (CLLD) framework, https://clld.org.

[4] https://software.sil.org/cog/.

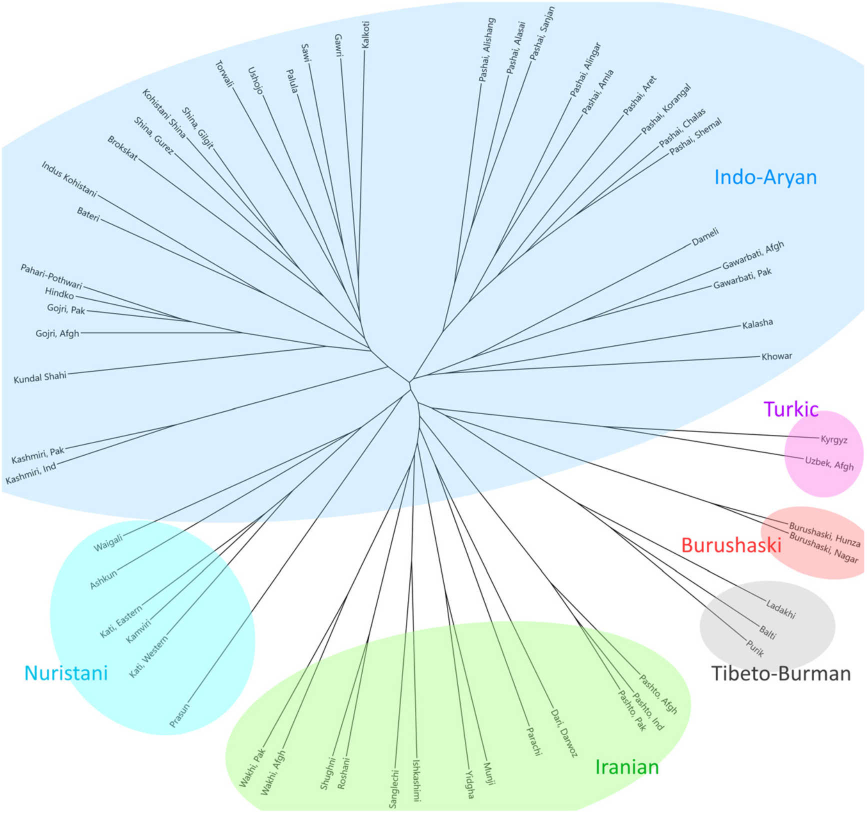


Figure 3: Automated clustering with Cog version 1.3.4.10016 (lexical neighbor-joining tree based on 100 basic vocabulary items).

varieties each cluster within their own phylogenetic group, although Nuristani Prasun is an outlier in relation to the other five Nuristani varieties. That deviance may be due to more radical phonological development, as compared to the rest of the Nuristani varieties, thus obscuring cognate identification, but may in itself be attributable to substratal effects, as suggested by Nelson (1986: 65). The close proximity displayed between some of the Iranian languages is fully expected, as the one between the three Pashto varieties; the one between the two Wakhi varieties; the one between Shughni and Rushani, considered belonging to the same sub-group (Wendtland 2009: 172–173); the one between Sanglechi and Ishkashimi, sometimes treated as dialects of the same language (Payne 1989: 419); and the one between Yidgha and Munji, two closely related, but gradually diverging, varieties (Skjaervø 1989: 411). The appearance of Darwazi, essentially a local Badakhshani variety of West Iranian Dari (Beck and Beyer 2013), alongside Parachi, an East

Iranian language, is rather more unexpected, but may at least partly be attributed to the advanced Persianization of Parachi that Kieffer describes (2009: 716–717). In a general sense, this analysis, once again, challenges "Pamiri" as an intermediate unit of phylogenetic classification.

The languages classified as Indo-Aryan present us with a much more disparate picture. They appear in two larger clusters that are at least as distant from each other as Nuristani is from either of them. The Indo-Aryan languages that cluster at the right hand side in the tree correspond to Bashir's (2003: 824–825) "Dardic" sub-groups named Pashai, Kunar and Chitral. In the cluster at the left hand side, we find languages belonging to the remaining "Dardic" sub-groups, Kohistani, Shina and Kashmiri, but the cluster also includes Indo-Aryan languages that have not traditionally been regarded as "Dardic", namely Hindko, Pahari-Pothwari and Gojri. Of these, the two former are members of a geographically very wide-spread continuum of Northwestern Indo-Aryan, with its main distribution in the lowlands of the Punjab, south of the HK; the latter is a Central Indo-Aryan variety belonging to a Gujarati-Rajasthani grouping whose main distribution is in the region between Delhi and the Arabian Sea in Western India. This analysis once again puts the validity of a distinct "Dardic" phylogenetic grouping within Indo-Aryan in some doubt. A further observation to be made is that the earlier recognized subgroups Kohistani and Shina do not form two identifiable sub-clusters, but are instead intertwined with one another. The most convincing and distinctive sub-clusters in this analysis that can be tied to previous lower-level classification within Indo-Aryan are those of Kashmiri and Pashai.

However interesting the questions may be that this analysis raises, it should be emphasized that this in no way constitutes a classificatory statement, especially regarding lower-level groupings. The methodology is simply too crude, and any finer-grained re-classification will have to take e.g. regular sound changes into account, and differentiate, to the extent possible, between inherited and loan vocabulary for each variety under investigation. What it does do is confirm the overall phylogenetic entities in broad strokes, thus validating the data, while also problematizing and putting into question the usefulness of a family tree model to reconstruct a historical-linguistic scenario in a region such as the HK. Let us therefore move on to the structural analysis, at the heart of the present study.

5 Structural analysis

5.1 Overall structural analysis

Structural proximity was measured next. This was based on the analysis of the presence (value = 1) versus absence (value = 0) of 80 structural properties (see

Appendix for details on features and values). As explained in Section 3, a wide variety of linguistic properties were represented, roughly related to five domains that were given equal weight. Distance is simply calculated as the number of shared feature values between any two of the sample varieties. SplitsTree version 4.14.6 (Huson and Bryant 2006) produces a NeighborNet visualization of these distance relations as displayed in Figure 4. While it is not entirely clear how to interpret this, it is obvious that the overall structural similarities between the sample varieties are of a very different kind as compared to the shared cognacy discussed in the previous section. The lexical analysis largely lines up with established phylogenetic classification, whereas the structural analysis, at least partly, seems to reflect entirely different types of affinities between the sample languages.

The distinctiveness of Sino-Tibetan is still clearly discernable, but the Turkic varieties now cluster with most of the Iranian languages, and Burushaski with a subset of Indo-Aryan. Indo-Aryan is itself split up into several relatively tight clusters, but apart from a cluster exclusively made up of Pashai varieties, and another outlier consisting of the two Chitral Group languages Khowar and Kalasha, the other similarity clusters do not correspond to any established Indo-Aryan phylogenetic units. The Pashto varieties line up with Indo-Aryan languages as their closest neighbours, while appearing to be maximally different from the rest of the Iranian sample languages. Among the Nuristani languages, Prasun is an apparent outlier.

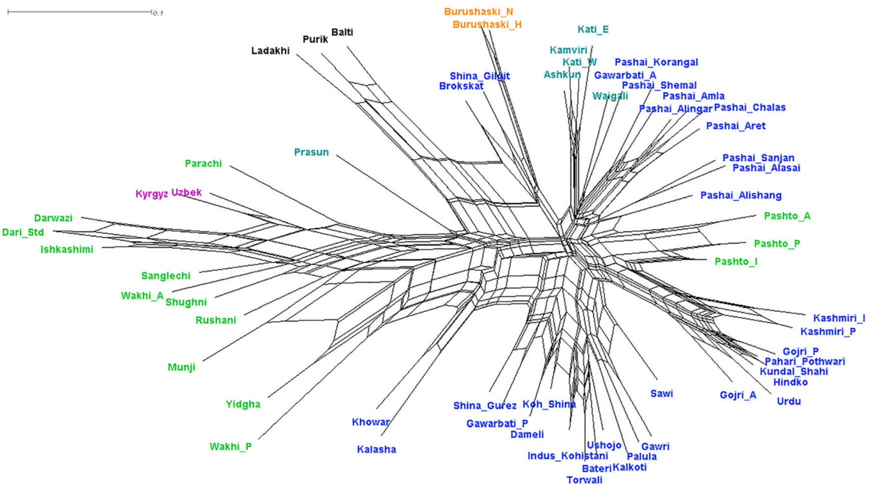


Figure 4: NeighborNet visualization of all 80 structural features generated by SplitsTree version 4.16.1.

We will have reasons to return to overall structural proximity and its possible relation to geographical proximity in Section 5.3.4, but for now it suffices to conclude that phylogenetic identity is only weakly reflected in structural similarity across the sample.

5.2 Domain by domain structural analysis

When each of the five structural domains is analyzed one at a time, a more differentiated picture emerges. Starting with word order properties, the region as a whole is largely homogeneous, but for a weak tendency for a few of the Iranian languages of the far northwest to deviate from the overall patterns. The latter languages use for instance prepositions rather than postpositions, and display the predominant order Object argument followed by an Oblique NP, while the opposite order is the more common in the rest of the HK. As for clause structure properties, the results are difficult to interpret and generalize over. The clearest clustering, however, appears to be sub-areal, setting off the northwest of HK, including the "Pamir" Iranian languages, the Turkic varieties, the adjacent Indo-Aryan languages Kalasha and Khowar, and Nuristani Prasun. This clustering is e.g. reflected in accusativity, such as verb agreement with A (i.e. the transitive subject) and accusative alignment of NP arguments (i.e. differential marking of the direct object), which is predominant in the languages of this subregion. Properties related to grammatical categories display a relatively high degree of homogeneity across the entire region, with the exception of Iranian – apart from Pashto, which in many ways deviate from what otherwise appears to be typical of languages in the HK.

As for lexical structure or the domain of lexico-semantic organization, the picture of a substantial core emerges, including languages of the central areas of the HK. Regardless of phylogenetic identity, Indo-Aryan, Nuristani, Iranian, Burushaski and Sino-Tibetan alike share features vis-à-vis languages spoken at the fringes of the HK, i.e. languages of the far northwest, the westernmost Indo-Aryan languages as well as the southernmost languages, including Iranian Pashto and a few Indo-Aryan languages. In regard to the phonological domain, a similar core emerges, although slightly more restricted than the one displayed for the lexico-semantic domain. This core includes a number of Indo-Aryan languages, Burushaski and Nuristani Ashkun.

It appears that the domains of word order, clause structure and grammatical categories display a similar type of clustering, and the lexico-semantic and phonological domains display another, rather different one. Therefore, by collapsing the former three into a more general morphosyntactic set of 48 features, and in this case using SplitsTree version 4.16.1 to apply UPGMA, a clustering algorithm whose output emphasizes deep splits to a larger extent than the NeighborNet algorithm, we get the tree representation in Figure 5. According to this representation, a smaller part of the sample (the one to the right in the tree), including 15 sample varieties, forms a distinct cluster, setting it apart from the majority of the HK languages (to the left). Both clusters are phylogenetically heterogeneous, although there is an Iranian predominance in the smaller one. This smaller cluster is clearly areal in nature, as these language communities exclusively belong in the northwestern part of the region under study, mainly in Afghanistan's Badakhshan province and adjacent areas of Pakistan (and Tajikistan). It is also worth noting that the two languages of wider communication, Urdu and Dari, are part of a cluster each, Urdu of the larger cluster and Dari of the smaller, northwestern, one.

When the 32 lexical and phonological features are similarly collapsed, we get the tree in Figure 6. In this case, too, it results in two distinct clusters. Once again, the two clusters are phylogenetically heterogeneous, but are of different compositions compared to the ones in Figure 5. The slightly larger one (to the left in the tree) includes all the Sino-Tibetan, all the Nuristani as well as both Burushaski varieties in the sample, in addition to the majority of the Indo-Aryan and a few Iranian varieties. The smaller one (to the right) includes the two Turkic varieties, most of the Iranian and some of the Indo-Aryan varieties. This clustering is also clearly areal in nature, but this time it takes the form of a core, whose member are all the languages spoken in the central parts of the HK region, versus the region's peripheries, including languages of the far northwest (the Turkic varieties and some of the Iranian languages spoken far from the centre of the HK), those in the

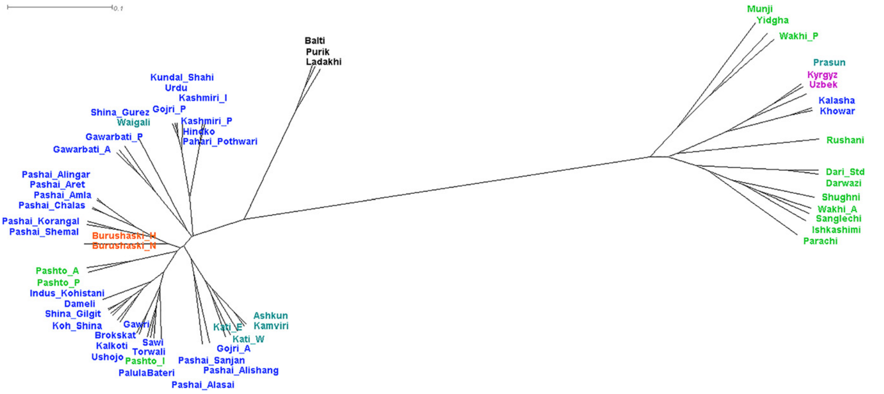


Figure 5: UPGMA tree based on 48 morphosyntactic features generated by SplitsTree version 4.16.1.

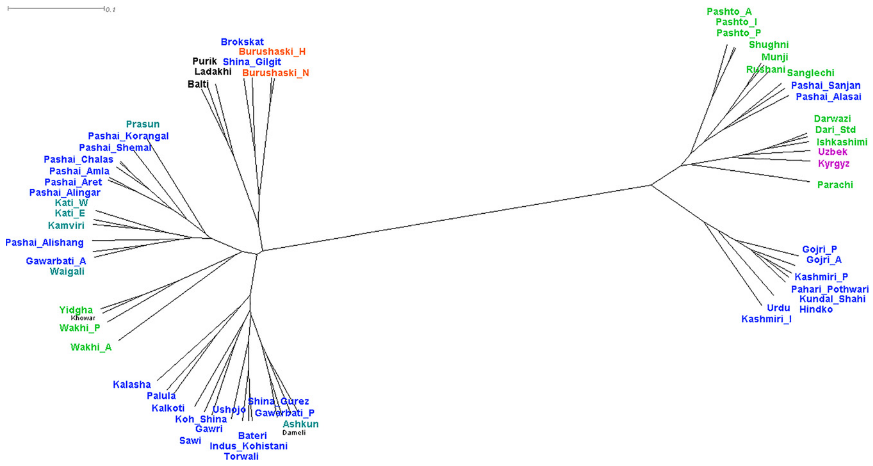


Figure 6: UPGMA tree based on 32 phonological and lexico-semantic features generated by SplitsTree version 4.16.1.

western outskirts (Iranian Parachi and the two westernmost Indo-Aryan Pashai varieties) and languages in the south or southeast, of which many belong to continua or have an origin linked to regions far outside the scope of the HK (such as Iranian Pashto and Indo-Aryan Gojri and Hindko). The languages of wider communication, Urdu and Dari, both belong in this peripheral cluster. It is worth noting that a few of the languages that clustered with the "deviant" northwest in regard to morphosyntax are part of the HK "core" in regard to phonology and lexico-semantics: Nuristani Prasun, Indo-Aryan Khowar and Kalasha, and Iranian Yidgha and Wakhi.

We will have reasons to return to these particular areal tendencies in Sections 5.3.3 and 5.3.4 as well as in the concluding section, but for now it suffices to conclude that different subsystems of language show different types of clustering, and that certain contact patterns therefore seem to be more apparent in some linguistic domains than in others.

5.3 Feature distribution and geographical correlations

In this sub-section, the focus will be on the distribution of individual, or a few closely related, features. Some significant distributional patterns observed are the following: a north versus south distribution, an east versus west distribution, a core versus periphery distribution, and finally patterns that are either micro-areal or macro-areal.

5.3.1 North versus south

The presence (or rather possibility) of encoding predicate nominals without an overt copula is one of the clause structure features that was investigated. This is for instance a possibility in Indo-Aryan Kalasha, as in example (1), and in Iranian Rushani, as in (2). Both of these exemplify equation, in which the two noun phrases have one and the same referent, and yet lack an element that explicitly links the copula subject with the copula complement.[5]

(1) Kalasha (Indo-Aryan) [kls]

ɕ-iːa	*motɕ*	*iɕkaːri*
EMPH-PROX.NOM.SG	man	hunter

‘This man is a hunter.’ (KLS-Val-LR:070)

(2) Rushani (Iranian) [sgh(r)]

jim	*tʃuruk*	*ɣiːwgar*
DEF.PROX	man	hunter

‘This man is a hunter.’ (SGHr-Val-ZB:070)

While this is a typical feature of Turkic, the same absence of a copula in present tense equational sentences has previously been reported for a few other individual languages in HK (Baart 1999: 118–122, 2009: 841–842, 2016: 267; Liljegren 2016: 303–306). It should be noted that this feature is treated somewhat differently here than the corresponding one in the WALS database (Stassen 2013). While in the WALS treatment, the two values are defined as “zero copula is possible” versus “zero copula is impossible”, in the present treatment, only the presence of zero copula in the data sets available has been coded as feature value = 1, while a feature value = 0 merely indicates its non-occurrence in the data, without making a statement about its impossibility per se. The assumption, based on observations of the phenomenon in individual languages, is that it is typical but not necessarily obligatory. The presence of zero copula has been detected in 16 of the sample varieties (Table 3).

While it is a minority feature in the region, it nevertheless appears to have a geographical distribution in the north, while it is absent in almost all of the varieties in the south (Figure 7).

[5] Outside the standard abbreviations of the Leipzig Glossing rules: EMPH = emphatic; REM = remote.

Table 3: Zero copula for predicate nominals.

Feature value	# of varieties displaying it	%
Present	16	27
Absent	42	71
Indeterminate	1	2

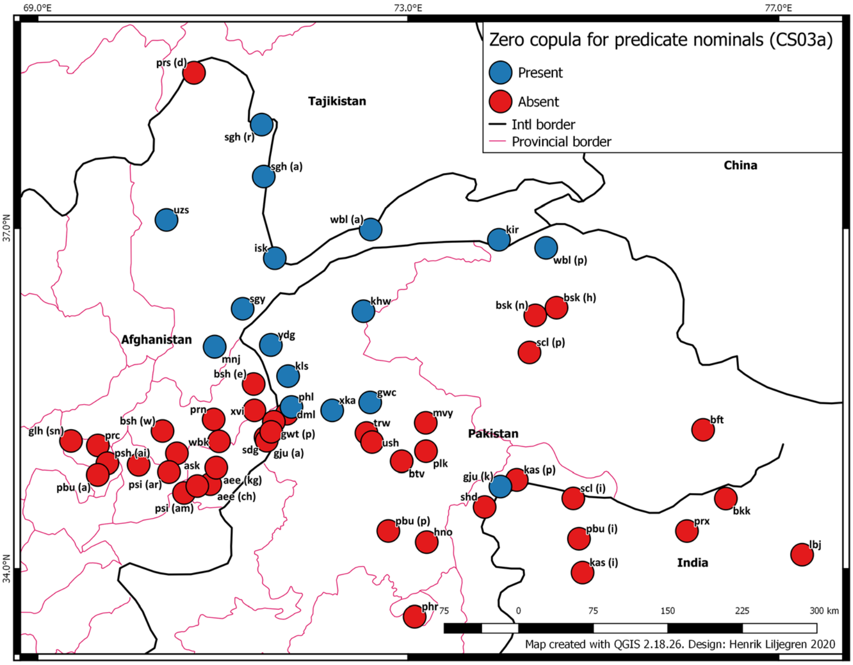


Figure 7: Zero copula for predicate nominals in the Hindu Kush.

Another feature with a north versus south distribution is the presence of uvular consonants. That a voiceless uvular plosive contrasts with a voiceless velar plosive is illustrated with near-minimal pairs in Burushaski of Nagar (3) and in Turkic Uzbek (4).

(3) Burushaski, Nagar (isolate) [bsk(n)]

/ʈak/	‘tied’	(BSKn-Val-SD:040)
/taq/	‘broken’	(BSKn-Val-SD:025)

(4) Uzbek (Turkic) [uzs]

/qiz/	‘daughter’	(UZS-Kin-WD:005)
/kʉz/	‘eye’	(UZS-40list-WD:009)

This velar versus post-velar contrast appears to be original in Burushaski (as well as in some other relic languages of North Asia), in which the contrast is upheld for unaspirated (k vs q) as well as aspirated (k^h vs q^h) plosives (see Table 10), whereas in Turkic, it has developed from allophonic to phonemic (Tikkanen 2008: 253–254). In the HK sample, the presence of uvular sounds is a majority feature (see Table 4), found in 33 of the varieties, predominantly in the northern half of the region. It should be noted, however, that presence does not necessarily imply a phonological contrast such as the one noted for Uzbek and Burushaski.

Although these sounds occur widely in Iranian, Indo-Aryan as well as in Sino-Tibetan in the HK region, it is a relatively recent innovation. In many of these languages it is not strongly established and is mainly associated with loan vocabulary of Perso-Arabic origin. It is not unusual that such sounds occur in free variation with velar sounds (e.g. q realized as x or k) and that people exposed to languages such as Arabic to a greater extent make a conscious attempt at a post-velar or uvular pronunciation compared to people lacking such exposure. An obvious exception to this tendency among the Indo-Aryan languages is Khowar, the Indo-Aryan language with the northernmost location. Contrasts between uvular and velar plosives as well as between uvular plosives and velar/post-velar fricatives are consistently upheld. It is also obvious that uvular plosives as well as velar/post-velar fricatives are part not only of loan vocabulary but occur frequently in basic vocabulary in that language (Bashir 2007: 217; Liljegren and Khan 2016: 222), as is illustrated in (5). Similar occurrences of indigenous vocabulary with uvular sounds, in addition to their occurrence in Arabic-derived loans, have been noted for e.g. Indo-Aryan Shina of Kohistan and Gawri (Baart 1997: 13, 30; Schmidt and Kohistani 2008: 32–33).

Table 4: Uvulars.

Feature value	# of varieties displaying it	%
Present	33	56
Absent	25	42
Indeterminate	1	1

(5)	Khowar (Indo-Aryan) [khw]		
	/ɖɑq/	‘boy’	(KHW-Ipa-AA:085)
	/ɖɔk/	‘carried on back’	(KHW-Ipa-AA:096)
	/mɔx/	‘face’	(KHW-Ipa-AA:082)

A property that belongs in the domain of lexico-semantic structure is the sequential order of elements in complex cardinal numerals, i.e. numerals that combine one (or more) multiplicational bases with a single lower numeral to express numerals 11–99. This typology is not to be confused with a differentiation between the use of decimal and vigesimal bases, respectively (a feature that will be discussed in Section 5.3.3). Here, too, a north versus south correlation is clearly noticeable. Some higher numerals follow the pattern BASE + n, meaning that e.g. the multiplicational base 20 in ‘22’ precedes n, the single lower numeral, while other languages follow the reverse order, i.e. n + BASE. In the HK sample, some languages consistently display one or the other pattern. As shown in Table 5, numerals in Indo-Aryan Kashmiri are constructed along the n + BASE pattern, whereas the corresponding numerals in Turkic Uzbek follow the pattern BASE + n. However, some languages in the sample construct numerals in different ways depending on the particular numeral base that is used. In Nuristani Ashkun, the numerals 11–19 follow one pattern, namely n + BASE$_{10}$, while the numerals 21–29 display the reverse sequence, BASE$_{20}$ + n. It has therefore been necessary to treat the compositions involving bases of 10 as a feature separate from compositions involving bases of 20 in order to arrive at a more comprehensive typology.

As can be seen in Table 6, numeral composition with the base 10 follows the pattern BASE + n in 14 of the sample languages, i.e. a minority, whereas the feature value is absent in 45 of them, which in this case means that they follow the reverse order (n + BASE$_{10}$).

Numeral composition with the base 20 that follows the pattern BASE + n is a majority pattern (Table 7), followed in 40 of the sample languages, with 17 using

Table 5: Numeral composition with bases 10 and 20.

		‘6’	**‘16’**	**‘26’**
n + BASE	Pashto, P (Iranian)	***ʃpag***	***ʃpaːɽəs***	***ʃpagiʃ***
	Kashmiri, I (Indo-Aryan)	*ɕʲe*	*ɕura*	*ɕatwu*
n + BASE$_{10}$/BASE$_{20}$ + **n**	Ashkun (Nuristani)	***ʂu***	***ʂu***_ɽis_	*wiɕaː* ***ʂu***
	Brokskat (Indo-Aryan)	***ʂa***	***ʂø***_beɕ_	*biʑi* ***ʂa***
BASE + **n**	Wakhi, A (Iranian)	***ɕaːd***	*ðas* ***ɕaːd***	*iːbistet* ***ɕaːd***
	Uzbek (Turkic)	***alte***	*won* ***alte***	*jegirma* ***alte***

Table 6: Numeral composition BASE$_{10}$ + n.

Feature value	# of varieties displaying it	%
Present	14	24
Absent	45	76
Indeterminate	0	0

the reverse n + BASE$_{20}$ pattern, and for the remaining two we cannot discern from the data which pattern is used.

In order to capture the total picture, the geographical distribution of the two features have been combined in the map displayed in Figure 8. What the map shows is that the consistent sequential order n + BASE occurs exclusively in the southern half of the region, primarily in Indo-Aryan but also in Iranian Pashto and in Nuristani Prasun. The consistent order BASE + n occurs in the northern part of the region (in Turkic, Burushaski, and in the Indo-Aryan and Iranian languages of that sub-region) as well as in the three Sino-Tibetan varieties. The order n + BASE$_{10}$ and BASE$_{20}$ + n can be regarded as belonging to an overlapping zone, as it has a much wider distribution and occurs both in the northern and the southern parts of the region. This presents an overall picture of the order BASE + n as a northern feature that gradually fades out toward the south. Bashir (2016: 265) has previously pointed out the BASE$_{10}$ + n pattern as a strong indicator of areality in the Indo-Aryan languages that have it (Kalasha and Khowar), clearly contrasting with the inherited Indo-Aryan n + BASE$_{10}$ pattern.

A feature whose presence is geographically correlated with the south rather than with the north is verb agreement with the P argument (usually the same as a direct object in a transitive clauses), occurring in certain verbal categories with past or perfective reference. In the sample it is represented by 26 of the languages, and its absence characterizes 32 of them (Table 8). Apart from its geographical correlations, the presence of the feature is the result of well-documented historical developments in older stages of Indo-Iranian. In our sample it is found in a number

Table 7: Numeral composition BASE$_{20}$ + n.

Feature value	# of varieties displaying it	%
Present	40	68
Absent	17	29
Indeterminate	2	3

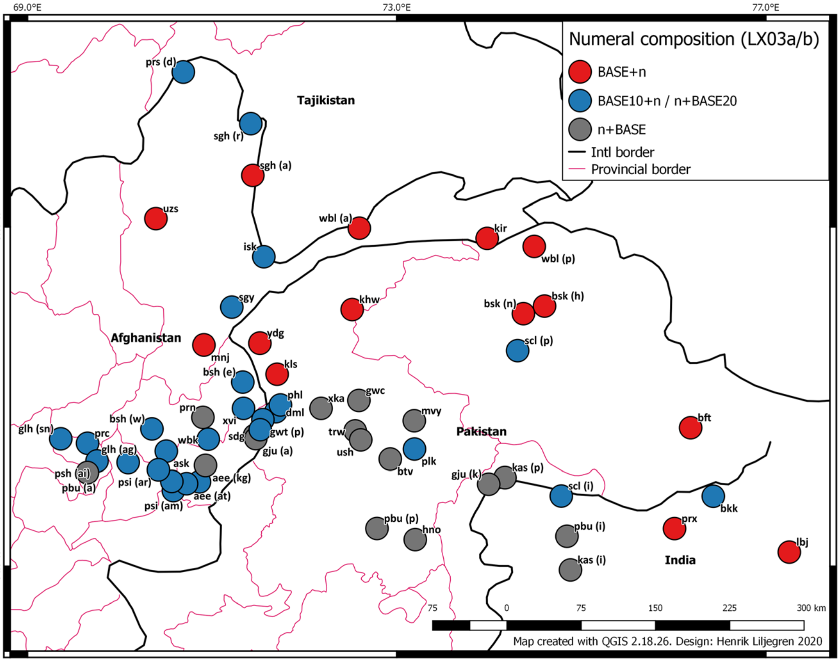


Figure 8: Numeral composition BASE + n in the Hindu Kush.

of Indo-Aryan languages, in most of the Nuristani languages, in Iranian Pashto as well as in Burushaski. It is entirely absent in the Turkic and Sino-Tibetan varieties.

Such P argument agreement is associated with one type of split ergativity that is common in large parts of South Asia, here illustrated by Pashto as spoken in Afghanistan (6). While the direct object 'scream' occurs in the direct (or default) case, the subject of this transitive clause, 'man', receives oblique case marking. The past tense-inflected verb agrees not with the subject but, in gender and number, with the direct object. Transitive clauses in present tense, however, follows a consistent nominative-accusative pattern. While the ergative pattern is

Table 8: Patient agreement in verbal past categories.

Feature value	# of varieties displaying it	%
Present	26	44
Absent	32	54
Indeterminate	1	2

linked to past tense in Pashto, it is instead linked to perfective aspect in many Indo-Aryan languages.

(6) Pashto, Afghan (Iranian) [pbu(a)]

saɽiː	***tʃiʁa***	*kɽ-əl-**a***
man.OBL	scream(F)	do.PFV.PST-PST-**3.F.SG**
A	P	

‘The man screamed [lit. made scream].’ (PBUa-Val-KO:058)

The absence of P agreement in the majority of the languages (including many Indo-Aryan and all Iranian except Pashto), however, is not straightforwardly related to the absence of ergative case marking. While it is true for Nuristani Prasun that a nominative-accusative agreement pattern combines with nominative-accusative case marking, as can be seen in (7), nominative-accusative agreement in Gurezi Shina is instead combined with ergative case-marking, as can be seen in (8).

(7) Prasun (Nuristani) [prn]

anzuː	*kijur-an*	*aːtɕaːwo**-mə***
1SG.NOM	child-PL.OBL	help.PST**-1.SG**
A	P	

‘I helped the boys.’ (PRN-Val-SM:015)

(8) Gurezi Shina (Indo-Aryan) [scl(i)]

beːs	*tɕuːɳoː-ɽ*	*kitaːb-eː*	*di-eːs*
1.M.PL.ERG	child-PL.DAT	book-PL	give-PFV.**1.M.PL**
A		P	

‘We gave the books to the children.’ (SCLi-Val-SA:036)

A related issue, not to be further discussed here, is the simultaneous indexing of A and P arguments, occurring in what appears to be three separate parts of the region: in some of the Indo-Aryan Pashai varieties in the southwest, in Indo-Aryan Kashmiri in the southeast, and in the isolate Burushaski in the northeast.

A completely different property that shows a geographical distribution quite similar to that of P agreement is a lexico-semantic one, namely the construction of compounds referring to one’s parents. In the sample, three distinct types of ‘parent’ words occur: a) lexemes without any obvious relation to words referring to ‘mother’ or ‘father’, b) compounds structured sequentially as mother + father, and c) compounds structured sequentially as father + mother. Typically, the two roots in the b) and c) types, are juxtaposed, but explicitly conjoining elements occur, too. Examples are displayed in Table 9. As can be seen, the M-F compounds are restricted to the three Indo-Iranian phyla, while the F-M compounds occur in languages of all six phyla.

Table 9: Mother-Father compounds.

M-F compound		F-M compound	
oj-bəw	Aret Pashai (Indo-Aryan)	*daːda-ʑe-aːja*	Kalasha (Indo-Aryan)
maː-wo-baw	Parachi (Iranian)	*taːt-ə-naːn*	Sanglechi (Iranian)
jej-tati	Waigali (Nuristani)	*jeː-nan*	Prasun (Nuristani)
		ata-aŋo	Balti (Sino-Tibetan)
		aːta-aːnæ	Uzbek (Turkic)
		au-ami	Hunza Burushaski

The M-F compounds largely occur in the south or in the southwest, while the F-M compounds primarily occur in a belt stretching from the northwest across the region to the southeast, interspersed with languages using other lexical expressions (Figure 9).

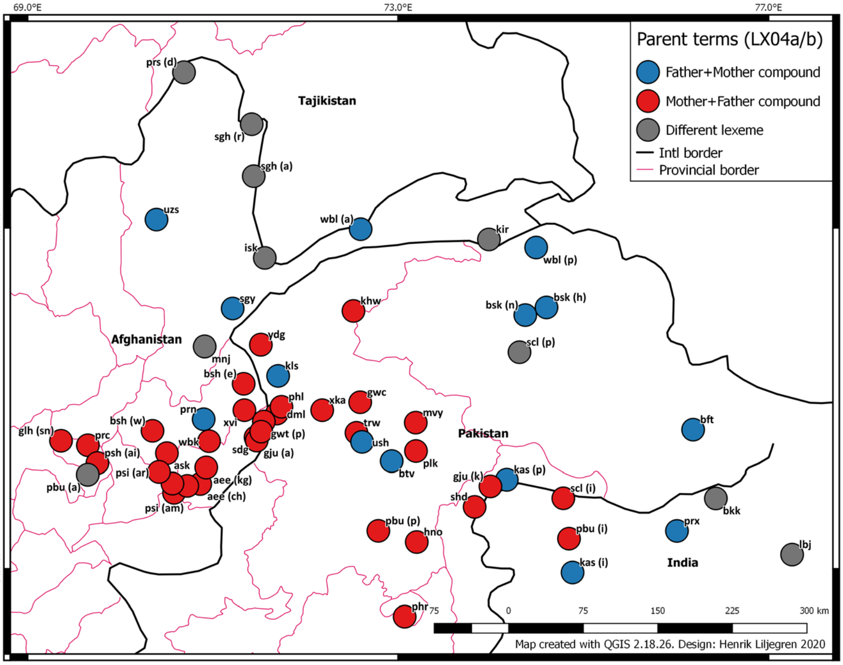


Figure 9: Parent words in the Hindu Kush.

5.3.2 East versus west

A phonological property that was studied and coded, despite analytical challenges, was contrast in aspiration. This phonemic contrast is an inherited property in Indo-Aryan, in Burushaski (Toporov 1971: 111) and possibly but not necessarily in Sino-Tibetan. In the latter case, it may have been allophonic in earlier stages, but later phonemicized, at least partly because of Indo-Aryan vocabulary being incorporated into Tibetan (Hill 2007: 489). The contrast was absent in earlier stages of Turkic, Iranian as well as Nuristani, although in regard to the latter two, voiced aspiration was a feature of the Indo-Iranian precursor common to Indo-Aryan, Nuristani and Iranian (Burrow 1973: 67–73). We leave contrast in voiced aspiration aside, as that poses even more interpretational challenges in the relatively few HK languages (almost exclusively Indo-Aryan, and always in combination with voiceless aspiration) that seem to uphold such a contrast. Even so, a contrast between voiceless aspirated and voiceless unaspirated consonants appears in as many as 30 of the sample varieties. These belong, as expected, to the three phyla in which it is, at least on some level, an inherited feature, i.e. Indo-Aryan, Burushaski and Sino-Tibetan. In all cases, it is a feature of one or more subsets of consonants, primarily of plosives, but in some of the languages also of affricates, as in Burushaski: see the partial inventory in Table 10, and (9) which illustrates the recurring 3-way contrast: unaspirated voiceless versus aspirated voiceless versus voiced.

(9) Burushaski, Hunza (isolate) [bsk(h)]

/tap/	'leaf'	(BSKh-40list-SK:018)
/t^{h}ap/	'night'	(BSKh-40list-SK:024)
/dan/	'stone'	(BSKh-40list-SK:032)

The feature is absent in the languages belonging to the remaining three phyla, with two notable exceptions: Iranian Parachi, spoken in the southwesternmost part of the region, and Iranian Yidgha, spoken in western HK, adjacent to Indo-Aryan. In

Table 10: Plosives and affricates in Hunza Burushaski (Berger 1998: 13).

p	**t**		**ʈ**	**k**	**q**
p^h	t^h		ʈ^h	k^h	q^h
b	d		ɖ	g	
	ts	tɕ	ʈʂ		
	ts^h	tɕ^h	ʈʂ^h		
		dʑ	ɖʐ		

Parachi voiceless as well as voiced aspiration occurs and appears to be contrasting. This, however, is not a preserved Iranian feature, nor is it contact-induced; rather, it has developed secondarily (Kieffer 2009: 694–695). In Yidgha, the evidence is not conclusive and will have to be further investigated, but tentatively both clearly aspirated and clearly unaspirated plosives occur in the present material, also confirmed by the consultant's own (Arabic-based) written representation. If this is a new development (not noted in closely related Munji, spoken further to the west), it may be attributable to Indo-Aryan Khowar influence, the locally dominant language, possibly reinforced by Urdu as the predominant medium of education. The map in Figure 10 displays the geographical distribution of voiceless aspiration contrast in the eastern part of HK, and its almost complete absence in the west (and north). The Pashto varieties of Pakistan and Indian Kashmir represent relatively recent speaker migrations.

What remains of interest from an areal perspective is the relatively recent loss of erstwhile aspiration contrast. This has taken place, and may still be under way, in a number of individual Indo-Aryan varieties, all of which are spoken in the westernmost part of the region, and primarily on Afghan territory. This

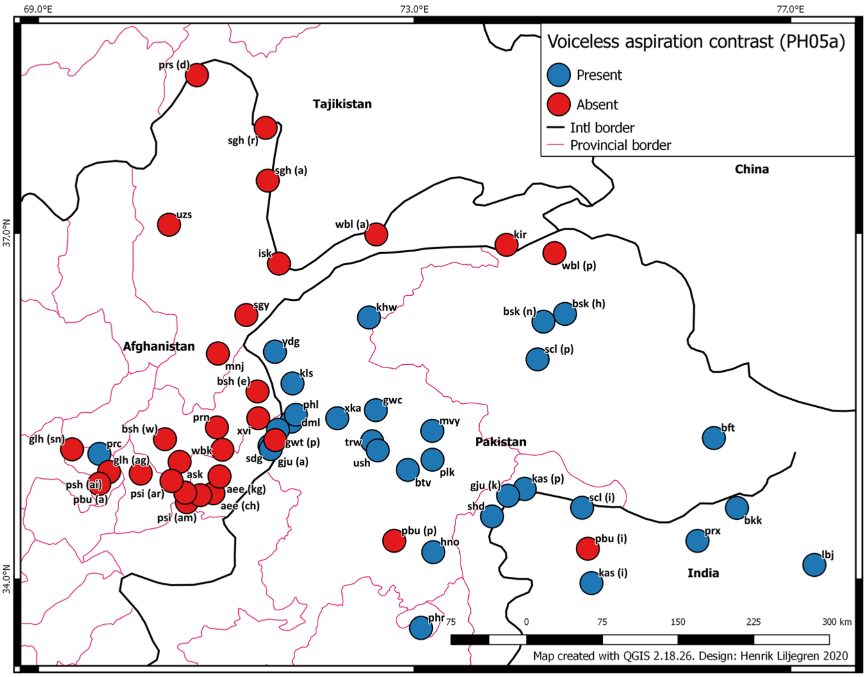


Figure 10: Voiceless aspiration contrasts in the Hindu Kush.

development, as far as the Pashai varieties are concerned, was noted already by Morgenstierne, who still (in the first half of the 20th century) noted aspiration in some cases, and in some of the varieties, but remarked that that it was "vacillating" (1967: 28–29), and again by Buddruss who (by the mid-20th century) was not able to determine any kind of aspiration contrast with certainty (1959: 5), and finally by Lehr, who studied a southeastern variety spoken in Darra-i-Nur and concluded that there is no contemporary evidence of an aspiration contrast (Lehr 2014: 12) – a finding confirmed, or at least not contradicted, by Lamuwal and Baker (2013) as well as by the present study. Indo-Aryan Gawarbati, a language spoken in the nearby Kunar valley, appears to display partial loss of aspiration. Morgenstierne (1950: 7–8) remarked that inherited voiced aspiration could still be detected in the oldest speaker generation, while voiceless aspiration remained contrastive, although some speakers replaced the voiceless aspirated bilabial with a fricative. Recent data, including the set obtained for this study, seems to suggest that Gawarbati as spoken on the Afghan side of the border is void of any aspiration contrast, while the variety spoken on the Pakistani side retains some voiceless aspiration contrasts, but is primarily limited to the velar and to a lesser extent the dental places of articulation, a finding that will need a more detailed study. Even for Indo-Aryan Gawri, spoken even further to the east, Baart (1997: 20) remarks that "aspiration seems to be losing its contrastiveness" or being replaced by tonal distinctions. Zoller (2005: 12–13) suggests that a gradual weaking of the aspiration contrast, as one moves westward, is due to substratal influence, an idea in line with Tikkanen's (1988: 308) earlier proposal that the complete loss of aspiration contrast in Nuristani, as well as the partial loss in Indo-Aryan in the adjacent areas, is an effect of a substratum common to the western parts of HK.

Another property that has a west versus east distribution is the presence of an initial polar question particle in the western part of the HK, as exemplified by Iranian Munji in (10), and an utterance-final question particle in the central and eastern part, as exemplified by Iranian Yidgha in (12). The latter is typically vocalic and cliticized to the utterance-final word. Interestingly, however, these two markers are not necessarily mutually exclusive of one another. In a few languages, both mechanisms may be at use, even in one and the same utterance, as can be seen in the Iranian Shughni example in (11).

(10) Munji (Iranian) [mnj]

oːjoː	*wə*	*tuːpə*	*ʃtə*
Q	REM.NOM.SG	ball	2.SG.POSS

'Is that ball yours?' (MNJ-Dem-DM:005a)

(11) Shughni (Iranian) [sgh(a)]
oːjoː *jaː* *tuːp* *tønd* = ***oː***
Q DEF.F ball 2.SG.POSS.PRED = **Q**
'Is that ball yours?' (SGHa-Dem-LL:005a)

(12) Yidgha (Iranian) [ydg]
ju *tɕaːɳɖuːl* *tukaːn* = ***a***
DIST.NOM.SG ball 2.SG.POSS.PRED = **Q**
'Is that ball yours?' (YDG-Dem-SZ:005a)

The initial polar question particle appears to be a relatively recent development in most of the languages that have it (including the Turkic varieties, as well as some of the Nuristani and Indo-Aryan varieties spoken in Afghanistan), and is most likely a superstratal "import" from Dari. The presence of the utterance-final question particle is probably of a much earlier date and seems to be a characterizing feature of the HK region itself (see Section 5.3.3), and its origin remains unknown (Morgenstierne 1938: 165).

A property of considerable antiquity is the use of possessive suffixes, particularly for inalienable possession. Although its scope and frequency differs from language to language, it is almost exclusively found in the western part of the region (in Iranian, Indo-Aryan as well as in Turkic, while absent in Nuristani), with or without a preceding free possessive pronoun, while the corresponding expression in the eastern parts only involve a preceding possessive pronoun, except for Burushaski, where a bound possessive prefix fulfills the same function as the possessive suffixes in the west. See Table 11 for examples of the different constructions used for 'my hand/arm' in the utterance 'My hand/arm is hurting'.

A final property worth mentioning in regard to a west versus east distribution is grammatical gender. This is an inherited feature of Indo-Aryan, Nuristani, Iranian and Burushaski. Subsequently, it is in those phyla that grammatical gender appears in the present sample, whereas it is entirely absent in Turkic and Sino-Tibetan. Most interesting for our purpose is therefore its retention, loss, or possible modifications in the four former phyla. Its realization in the Indo-Aryan languages

Table 11: Constructions used in inalienable possession, 'my hand/arm [is hurting]'.

Poss suffix		**Poss prefix or pronoun**	
*ast-**iːm***	Sanjan Pashai (Indo-Aryan)	***a**-ɕak*	Burushaski, Nagar (isolate)
*duːst-**um***	Parachi (Iranian)	***mẽː** tʰeːr*	Gawri (Indo-Aryan)
*quːl-**um***	Uzbek (Turkic)	***ŋa** lakpa*	Ladakhi (Sino-Tibetan)

of the HK region has already been addressed at length elsewhere (Liljegren 2019), but suffice it to summarize that Indo-Aryan in the eastern part of the region, like many other Indo-Aryan languages in the wider region, display a robust sex-based two-gender system, while the two northwesternmost languages Khowar and Kalasha have lost this type of gender differentiation altogether, while a differentiation between animate and inanimate has emerged in them. In yet another set of Indo-Aryan languages (primarily Pashai varieties), geographically all in the southwest, an animacy-differentiation has come to overlap with the inherited sex-based gender system. In Nuristani, a sex-based gender system, ultimately of the same origin as the Indo-Aryan, is retained in all of the sample varieties except for Prasun. Iranian displays a more varied picture. A sex-based system is retained in seven of the 13 sample varieties and is lost altogether in the other six. Burushaski, finally, is a four-gender system, essentially constituting a combined sex-based and animacy-based system.

5.3.3 Core versus periphery

The combined phonological and lexico-semantic multi-feature analysis in Section 5.2 identified the central parts, or the core of the HK region, as characterized by features not shared by the peripheries (and implicitly not by surrounding adjacent areas outside of the region). A few individual properties that exemplify this distribution will first be discussed below, followed by an attempt at testing whether membership in a putative HK contact area might be a matter of degree, with an increasing number of shared core features as one moves towards its geographical centre, rather than a linguistic area with clear boundaries.

A lexico-constructional pattern that was suggested as area-typical or area-defining in Koptjevskaja-Tamm and Liljegren (2017: 217) is the F = FB polysemy, i.e. the equation of one's father with one's father's brothers. Typically, an older brother of one's father is referred to as 'big father' and a younger as 'little father'. This pattern is not limited to a few individual languages, or even to one or the other linguistic phyla. Instead, it is attested in as many as 19 of the sample languages, representing four of the six phyla (viz. Indo-Aryan, Nuristani, Sino-Tibetan and Burushaski). A parallel, and obviously related, pattern is the M = MZ polysemy, equating mother with mother's sister. These two often appear in the same language, i.e. in the same kinship system, but that is not always the case. There is evidence of a M = MZ pattern in slightly fewer languages than those in which the F = FB pattern occurs. In our sample the M = MZ pattern has been verified in 13 of the 59 sample varieties. It was therefore decided to treat them as two separate features. An example of a kinship system in which both are expressed is that of Indo-Aryan Brokskat, as can be seen in Figure 11.

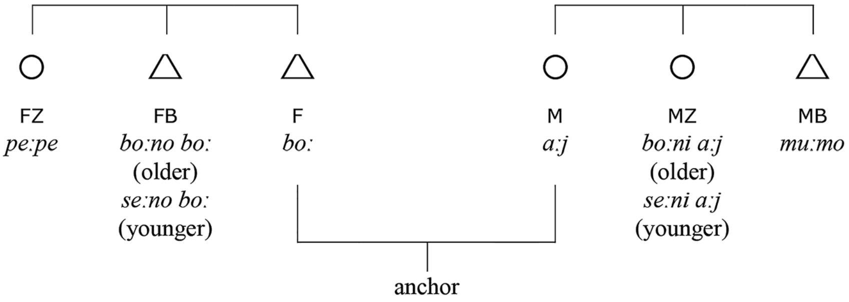


Figure 11: Terms for parents and their siblings, Brokskat (Indo-Aryan) [bkk].

Looking at the geographical distribution (Figure 12), the F = FB pattern occurs in a contiguous belt stretching through the inner, high altitude parts, of the region, from Nuristan in the west, through the northernmost parts of Pakistan, to Ladakh in the east. The M = MZ pattern, while attested in fewer languages, follows approximately the same trajectory.

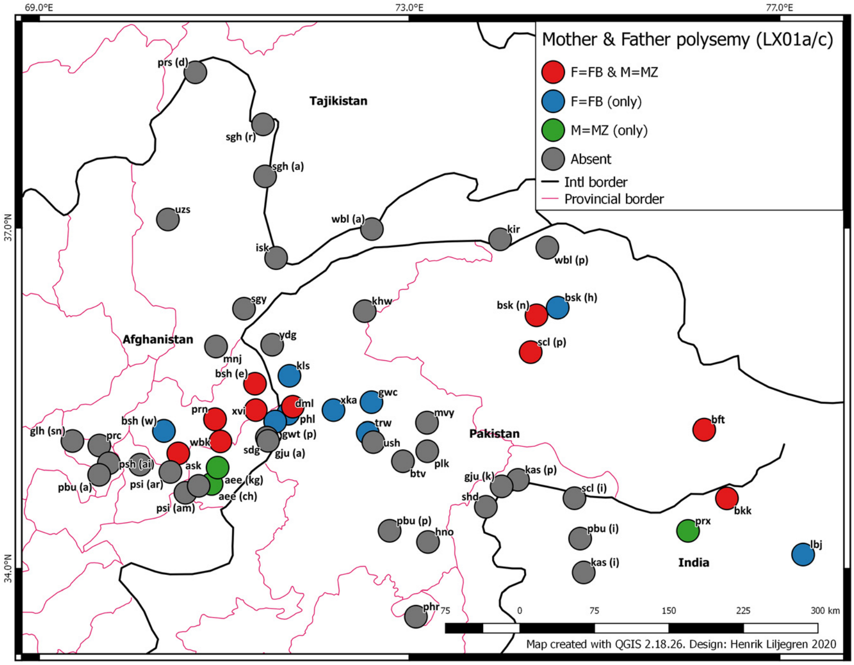


Figure 12: Polysemous terms for parents and their siblings.

Another, unrelated lexico-semantic feature that is typical of the HK region, but not of the surrounding regions of South or Central Asia, is numeral systems including a vigesimal multiplicational base (always in addition to a decimal one) appearing in the composition of the numerals 20, 40, 60 and 80. This is a majority feature in the HK, present in 38 of the sample varieties, absent only in 20, and indeterminate for one variety (Table 12). Apart from Turkic it is represented in all of the region's phyla. Its geographical distribution is reminiscent of the one described for the polysemy patterns above, but forms a thicker belt or crescent through the region, from west to east.

Two phonological features with a similar geographical distribution are the presence of retroflex affricates and the presence of retroflex fricatives. Retroflex affricates are part of the phoneme inventory of 24 of the HK sample languages, and retroflex fricatives of as many as 33 of them. In the present material, these two types of sounds are distinctly phonemic in languages of all the phyla except for Turkic and Sino-Tibetan. In the latter case, however, a phonetically retroflex fricative [ʂ] has been described by Zemp (2018: 30–33) as the voiceless counterpart of /r/ in Purik, and the retroflex affricates [ɖʐ] and [ʈʂ] have been described by Bielmeier (1985: 53) as sounds that at least phonetically occur in Balti. Therefore, this is a feature which would need to be looked at in further detail in those two languages in order to fully determine their phonological status. The parallelism between the phoneme inventories across many of the languages and even across phyla is quite striking (Nikolaev and Grossman 2018: 571–572; Zoller 2005: 14), particularly in respect to their complete tripartite dental–retroflex–postalveolar[6] contrast in combination with an almost complete tripartite plosive–affricate–fricative contrast, as schematically laid out in Table 13. Languages that exemplify this type of simultaneously affricate-dense and retroflex-dense inventory (with

Table 12: Vigesimal numeral base.

Feature value	# of varieties displaying it	%
Present	38	64
Absent	20	34
Indeterminate	1	2

[6] Postalveolar covers in this case a place of articulation that is variously described as alveolo-palatal, postalveolar or even palatal. These typically contrast with their laminal articulation vis-à-vis the apical articulation of retroflex sounds.

Table 13: Overlapping tripartite dental–retroflex–postalveolar and plosive–affricate–fricative sets.

t d	ʈ ɖ	
ts (ʣ)	ʈʂ (ɖʐ)	tɕ (dʑ)
s z	ʂ ʐ	ɕ ʑ

slight variations as far as voicing and aspiration contrasts are concerned) are – among a number of other languages – Indo-Aryan Khowar, Nuristani Prasun, Burushaski and Iranian Wakhi. In some of those language there is a tendency for the voiced fricatives and affricates to stand in an allophonic rather than fully contrastive relationship to one another. These inventories clearly contrast with those typical of Central/West Asia as well as South Asia beyond the HK region. The languages of the former typically have no retroflex sounds at all, and often nothing but postalveolar affricates. The languages of the latter area typically only have retroflex plosives, and affricates at one or two places of articulation.

The presence of an utterance-final polar question marker, as mentioned previously, is another diagnostic feature of the core HK region vis-à-vis the peripheries and surrounding areas. In HK, it occurs in languages of all six phylogenetic groups. Verbal alignment in the past or perfective realms is another such feature, which is not diagnostic to the same extent, but nevertheless singles out many of the Indo-Aryan languages of the core HK region as contrasting with the dominant Indo-Aryan languages outside of the HK. Of the HK sample languages, 28 display a nominative–accusative alignment pattern, including a number of languages that simultaneously display an ergative–absolutive pattern in their case marking.

In an attempt to test whether it makes sense to describe the HK contact area as consisting of a hard core with languages in its geographical centre that share a higher number of area-typical features than the languages at its peripheries do, the seven aforementioned core features (the presence of a vigesimal numeral base, F = FB polysemy, M = MZ polysemy, retroflex affricates, retroflex fricatives, a final question marker, and nominative-accusative verbal agreement) were taken into account, and each language was classified according to a 0 to 7 scale, by which a language displaying all seven HK-typical feature values was classified as 7 and a language lacking all of them as 0. It is quite obvious from the map in Figure 13 that an increasing number of such features (corresponding to increasingly darker shades) appear as one moves towards the region's geographical centre, suggesting a distinct core with its western axis situated in Nuristan and an eastern axis in the Burushaski-speaking area in the Hunza valley.

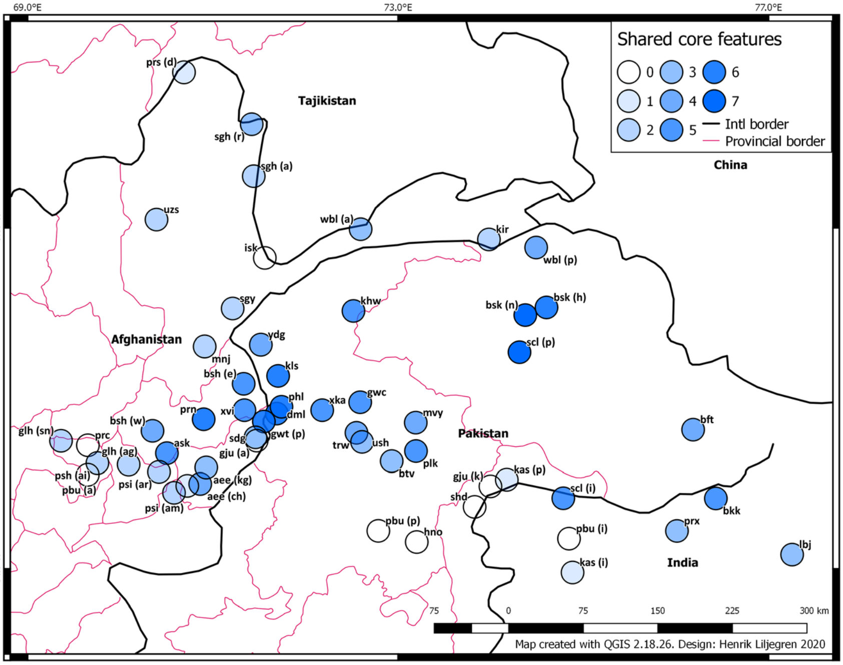


Figure 13: Number of shared core features in the Hindu Kush.

5.3.4 Macro-areas and micro-areas

In this final sub-section dealing with the geographical distribution of features, both a macro-areal and a micro-areal perspective will be taken. For the macro-areal perspective, the HK region, or a considerable part of it, will be looked upon as taking part in an even larger areal entity. For the micro-areal perspective, we will take note of significant, but considerably smaller and geographically defined clusters within the HK region. Those perspectives are in no way exclusive of one another. On the contrary, as pertinently pointed out by Masica (2001: 224), "the feature(s) define the area; the area […] does not *a priori* define the features," and an individual language variety may very well belong simultaneously to several different areas or contact zones, depending on geographical scope and what specific features are being investigated and compared across languages.

Starting from the top, HK is in its entirety part of what Masica terms an "Indo-Turanian" macro-area, a world region that includes most of Asia, except Southeast Asia and the Arabian peninsula (2001: 241–242). In the present data set, such inclusion is most clearly visible in features that display total or almost total homogeneity across HK. As previously mentioned (Section 5.2), this is primarily the case in the word order domain. Feature values, such as OV order (in 59 of 59 varieties), Adj N order (55/59), Dem N order (59/59), Gen N order (51/59), Poss N order (58/59), are bundled together, just like in large parts of the Indo-Turanian macro-area, and the exceptions to these values are few. Contrary to assumptions made in the early days of cross-linguistic studies of word order correlations, OV word order does not universally correlate with other head-final (i.e. modifier-noun) features; when they do, that head-dependent bundling is in itself area-specific (Dryer 1992), in this case typical of languages in large parts of Asia but not necessarily of OV languages in other parts of the world (Masica 2001: 241).

As far as a considerable number of features are concerned, however, a border between two major linguistic areas (both part of the larger Indo-Turanian "package"), appear to cut right through the HK region, namely one between a relatively distinct South Asian area and a less distinct Central (or possibly West) Asian area. This was foreshadowed in Section 5.2, and is clearly visible in Figure 5, in which the languages of the northwestern corner of HK display obvious feature bundling vis-à-vis the rest of the region, especially when aggregating all features of the morphosyntactic domains (word order, clause structure and grammatical categories). That this is not simply a matter of phylogenetic distribution that happens to coincide with geography is confirmed by the heterogeneity of both clusters. To test the significance of this division, eight diagnostic features were selected and applied to the clusters suggested by Figure 5, and the feature values for the 15 languages of the Central/West Asia cluster (CA) were compared with those of the 44 South Asian (SA) languages. The features are: the presence of a zero copula for predicate nominals, the order Object argument followed by an oblique argument, a unique P case, a unique A case, ergative alignment of nouns, sex-based gender, and verbal gender agreement. The proportions of feature value 1 (=present) are displayed in Table 14.

For most of the features, the great majority of languages in one of the clusters displays the feature, while the opposite is true of the other cluster, i.e. that a

Table 14: Proportions of the presence of eight morphosyntactic features within the Central/Asian (CA) and South Asian (SA) clusters, respectively.

	Prep	Zero cop (n)	Obj-obl	Unique P case	Unique A case	Erg align (n)	Sex gender	Gender agr
CA	**0.60**	**0.80**	**0.80**	**0.87**	0.20	0.20	0.03	0.07
SA	0.05	0.09	0.48	0.14	**0.86**	**0.91**	**0.93**	**0.89**

minority proportion displays it. The first four features are all typical of the CA cluster, while the remaining four are all typical of the SA cluster. While the presence versus absence of individual features is important, this obscures the gradual transitioning from one area to the other, visible in the presence of seemingly opposite features in one and the same language. While as many as 53 of the sample varieties have postpositions, for example, this does not necessarily mean the absence of prepositions. In fact, seven languages show a positive value for both, meaning that they have postpositions as well as prepositions; all of them are Iranian varieties. Some of those languages have relations that are encoded with postpositions and others by prepositions, while e.g. two of the Pashto varieties in the sample also use circumpositions for certain relations. Those Iranian languages can in a larger perspective be seen as transitional between a West Asian (and Persian-dominated) area characterized by the typologically less expected combination of OV word order and prepositions (Dryer 1992: 85), and the South and Central Asian areas characterized by the more common combination of OV and postpositions (Masica 2001: 241; Stilo 2009: 7). The idea of transit zones between larger contact areas has been applied to HK in its entirety by Tikkanen. He holds that at least certain, primarily phonological characteristics of HK as a convergence area are due to its dual membership in a Central Asian and a South Asian macro-area (2008: 258). Bashir (1996b: 177–178) offers a similar description of multiple membership in several non-exclusive, overlapping linguistic areas (South Asia and Central Asia in particular), with special reference to Indo-Aryan Khowar.

Returning to the proposals of Dahl and of Koptjevskaja-Tamm and Wälchli (referred to in Section 2), actual convergence, in which speaker communities stand in contact with one another and affect one another's speech in a tangible way, is primarily local, involving two, three, or perhaps a handful of languages. This pattern is clearly visible in the HK region, both when we apply feature aggregation and when we restrict ourselves to a detailed study of a few linguistic properties in particular sub-regions. If we include all the 80 structural features that were part of this study, and apply UPGMA as a clustering algorithm, six clearly discernable smaller contact zones or micro-areas emerge, as displayed in the map in Figure 14. One of these local contact zones diverges to a greater extent from the other five, the one labelled "Central Asian", including the Iranian languages of the Pamir-Wakhan region in the northernmost part of the HK region (but also Parachi) and the two Turkic varieties. This suggests that the other five micro-areas are all essentially part of the larger South Asian linguistic area. A West Hindu Kush micro-area, geographically centered in the provinces of Nuristan, Kunar, Laghman and Kapisa of northeastern Afghanistan, includes the nine Indo-Aryan Pashai varieties, five of the six Nuristani varieties as well as Iranian Pashto.[7] The three languages Khowar, Kalasha and Prasun (the two former Indo-Aryan and the latter Nuristani) form by themselves a North Hindu Kush micro-area, in some respects sharing properties with other South Asian languages, in other respects with Central Asian. An East Hindu Kush micro-area includes the two Burushaski varieties as well as 15 Indo-Aryan varieties spoken across the central highlands of the Hindu Kush, primarily in Pakistan's Gilgit-Baltistan province and the northernmost part of Khyber Pakhtunkhwa. The remaining two micro-areas are both part of the larger South Asian linguistic area, but are slightly more peripheral as far as Hindu Kush is concerned. What is here labelled "South Asian: Lowland" consists exclusively of Indo-Aryan varieties; apart from Kashmiri and Kundal Shahi, their closest linguistic relatives are Indo-Aryan languages spoken on the Indo-Gangetic plains southeast of the Hindu Kush region. The three Tibetan varieties form a small "Himalayan" cluster of its own, and as such form the northwesternmost extension of a distinctly Himalayan linguistic environment of northern India and Nepal.

A division of the Hindu Kush region into five so-called geo-cultural regions, i.e. regions that each displays a great deal of internal homogeneity in terms of cultural, social, political and religious identity and characteristics, is suggested by Cacopardo and Cacopardo (2001: 17), based on their extensive ethno-historical research. This non-linguistic division in fact lines up surprisingly well with our linguistic micro-areas outlined above, and will probably make even more sense once more detailed studies have been undertaken of linguistic properties characterizing the languages of such sub-regions. As mentioned already in the background in Section 2, the identification of a linguistic micro-area in the remote western central parts of the HK, for instance, coincides geographically in a significant way with the communities that were most recently incorporated into the Islamic world and for the longest time withstood conversion and the abandonment of their old world view and social practices. Another, equally significant, micro-area, is characterized by an "uncles and aunts" kin terminology (FB = MB/FZ = MZ), i.e. one different both in relation to the above-mentioned "mothers and fathers" terminology (M = MZ/F = FB) terminology (see Section 5.3.3) of the HK core, and the "full differentiation" terminology typical of the surrounding lowland regions (F ≠ FB ≠ MB); this one coincides largely with the North Hindu Kush contact zone, including language such as Indo-Aryan Khowar and Iranian Wakhi. Another set of linguistic properties whose distribution appears to coincide with the division of the HK region into micro-areas relates to alignment (Liljegren 2014: 162–167). If both

[7] The modern-day location of the Pashto varieties of Pakistan and India that were part of this study have not been plotted on the map, as these can be considered the result of more recent Pashtun expansion out of their historical heartland bordering on the southwestern part of the HK.

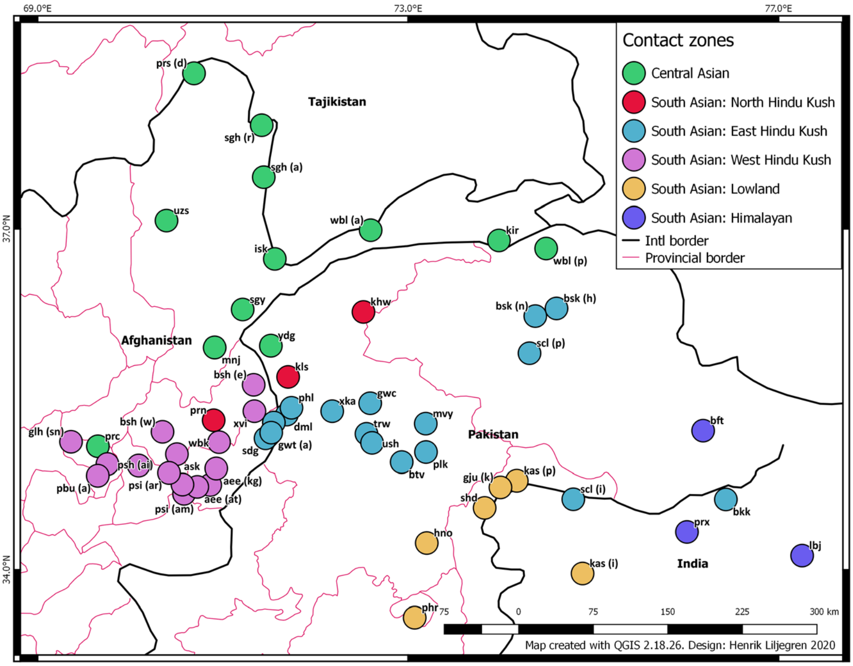


Figure 14: Micro-areas/contact zones in the Hindu Kush (based on 80 binary features clustering).

verbal and nominal alignment are taken into account, the northeast is characterized by a combination of verbal accusativity and nominal ergativity, often regardless of tense and aspect; the northwest is characterized by the absence (or radical weakening) of ergativity; and the south in characterized by nominal as well as verbal ergativity, albeit an ergativity that is largely limited to the realm of past or perfective categories. Some features or domains needing further research, but for which a micro-areal configuration is very likely to emerge, are those relating to discourse features, deictic systems, various lexicalization patterns and complex constructions.

6 Conclusions

Summing up the findings presented in the previous section, the largely diagnostic analysis of lexical proximity, or shared cognacy, shows a clustering in line with established classification into the six phylogenies represented in the sample, whereas the overall structural analysis reflects significantly different types of affinities between the languages, with obvious geographical correlations, and often crossing phylogenetic boundaries. It is particularly clear that the Burushaski varieties structurally cluster with a number of Indo-Aryan languages, and that the Turkic varieties cluster with the majority of the Iranian languages.

A more fine-grained structural analysis, in which the five separate structural domains are looked at separately, reveals a more differentiated picture as far as clustering is concerned. The domains of phonology and lexico-semantics display one pattern, while the morphosyntactic domains of word order, clause structure and grammatical categories, display a different one. In both cases, the feature distribution appears to be as much geographical as phylogenetic. As for the combined phonological-lexical domain, two larger affinity clusters can be discerned: a larger one comprising the languages in the central parts of the HK region, and another, slightly smaller one, including the languages in the peripheries. Both clusters are phylogenetically diverse. Also in the morphosyntactic domains, the languages belong to one of two affinity clusters: a considerably smaller one corresponding to the northwestern corner of the region, and another, larger one that is made up of the rest of the languages. Again, both clusters are multi-phylogenetic, but in this case, there is an Iranian dominance in the northwestern cluster and an Indo-Aryan dominance in the geographically more widely spread cluster.

In the analysis of individual structural features (or a few closely related features) and the distribution of each feature value (present or absent), a number of areal or geographical correlations stand out in particular. First, there is a north versus south distribution, which is reflected in a number of features belonging to several structural domains. Second, there is a west versus east distribution, partly as a secondary effect of phylogenetic distribution, partly reflecting a division along the international border between Afghanistan and Pakistan, each with a different national-level lingua franca that is dramatically reshaping and influencing the local languages within its realms – particularly, as it seems, syntactically. Third, there is a core versus periphery distribution, most aptly pointing to the HK region as consisting of a hard core, with a few languages in its geographical centre (also corresponding to the higher altitudes and its more remote parts) sharing a larger number of features, with a gradual decrease in the number of shared features, as one moves out toward the peripheries, until only a few or none of the features are represented in the languages spoken adjacent to lowland regions.

In addition, there are distributions that are even more local, pointing to the existence of at least six distinct micro-areas within the HK region, closely corresponding to smaller sub-regions with a shared historical, cultural and religious identity. Finally, there is another set of features, mostly related to the word order domain, for which the region as a whole, or a substantial part of it, is relatively homogeneous and thereby reflect its inclusion in even larger areal constellations, such as Masica's Indo-Turanian macro-area or a South Asian linguistic area. Tikkanen has already articulated ideas very similar to the conclusions of the present paper. He holds that the region that essentially overlaps with what I have here defined as the Hindu Kush–Karakorum "has several macro-areal features in common with other areal configurations, especially in South Asia. But in addition it has some micro-areal features diagnostic of itself or parts of itself." (2008: 254) He goes on to explain this particular cluster of micro- and macro-areal features as "the outcome of extended contacts, as supported by certain cultural isoglosses in this region" (2008: 258). These patterns, or layers, of areality are obviously of varying scope and historical depth. Each one of them can be linked to one or a few linguistic domains, and, together they give us an idea how the linguistic landscape of the Hindu Kush–Karakorum evolved over time.

In broad strokes it suggests the following time-line along which the region developed from a highly diversified linguistic landscape, phylogenetically as well as typologically, to today's relatively moderate level of phylogenetic diversity and an even more limited typological diversification: 1. Pre-Aryan equilibrium; 2. Aryan expansion; 3. Imperial encroachment; and 4. Post-colonial restructuring. The first period represents a distant past, before the main penetration of this mountain region by Indo-Aryan groups (and perhaps by Indo-Iranian in general), the latter taking place approximately 1500–1000 B.C. This ancient time period was most likely characterized by great phylogenetic diversity, perhaps akin to the Caucasus of today. Although diffusion may have affected the entire region, or substantial parts of it, changes must have been slow. It would be premature to exclude Indo-Iranian elements from having formed part of a gigantic pan-Himalayan accretion zone – a phylogenetic patchwork stretching through the Hindu Kush–Karakorum and continuing throughout the entire Himalayan high altitude region. However the progenitors of today's Burushaski and Sino-Tibetan languages were probably some of its more dominant components, along with languages or clusters of other stocks, whether Dravidian, Austroasiatic or of entirely unknown affiliations. Some shared lexico-semantic features, such as the dominance of vigesimal numeral bases, and some properties of kinship organization, are the only possible reflexes of language contact of this pre-Aryan era that were found in the present study.

The second period, from the second millennium B.C. to the initial arrival of Islam in 800–1200 A.D., is characterized by a marked decrease in deep diversity, as a result of populations shifting to politically more dominant, but initially demographically minor, Indo-Iranian (especially Indo-Aryan) languages spreading and growing in importance within the region – in the case of Indo-Aryan, expanding northward along several separate trajectories from the lowlands and lower slopes in the south. Modern-day Indo-Iranian languages that (structurally and/or lexically) most clearly show traces of early contacts with non-Indo-European languages of this period are Nuristani Prasun, Indo-Aryan Khowar and Iranian Wakhi, all located in the western and northern parts of the region, strongly suggesting the vital presence, at the beginning of that period, of a non-Indo-European/non-Burushaski substrate with its centre of gravity in today's Nuristan, and possibly another distinct substratal element centered somewhere in the Wakhan area. A similar substratal role was played by the ancestral language of today's Burushaski in the Gilgit area further to the east (with obvious traces in Shina varieties), and yet another one with a locus in Kashmir, underlying some of the particular developments of Kashmiri. Changes during this period are likely to have been rapid, producing area-wide innovation, further propelled by the aforementioned language shifts. The clearest reflexes of such area-wide diffusion and innovation can be seen in phonological features shared by Indo-Aryan, Nuristani, Iranian and Burushaski.

The third period, somewhat simplified, covers the time period from the arrival of Islam to the mid-20th century birth of the modern nation states. It is characterized by the spread of Islam as well as imperial claims on various parts of the region. Apart from a major influx of cultural innovation and lexical renewal, the general change rate probably slowed down, and to the extent that there were changes in linguistic structures, these were mainly a matter of genera-internal diversification and subareal diffusion. Specific local developments helped create smaller geo-cultural regions, such as the ones suggested above. As for clause structure features, e.g. those related to alignment patterns, their subareal clustering is largely tied to developments beyond the region itself. The final, modern-day period, its beginning phase largely coinciding with the withdrawal of British imperial aspirations, is characterized by restructuring on many different levels. What we observe is a dramatic decrease in diversity, both in number of speakers of the smaller languages, rapid language shifts, and structural, particularly syntactic, streamlining, largely driven by a few dominant and very large superstrata: Urdu, Dari and Pashto.

Acknowledgements: This work is part of the project Language contact and relatedness in the Hindukush region, supported by Vetenskapsrådet, the Swedish Research Council (421-2014-631). As its principal investigator, the author is responsible for the overall project implementation, but a large number of people and groups have contributed in crucial ways to the project. The main part of systematic data annotation and processing took place at Stockholm University. Noa Lange was employed as a research assistant, primarily working with IPA transcription, various types of data processing and ELAN annotation. As part of a project internship, Nina Knobloch fine-tuned data and prepared the output of the structural analysis. As part of the data collection process, collaborative elicitation workshops were arranged under the auspices of three institutions based in the target region: Forum for Language Initiatives (Islamabad, Pakistan), Samar (Kabul and Faizabad, Afghanistan) and the Department of Linguistics, University of Kashmir (Srinagar, India). The contributions of those institutions were primarily in providing logistic support, identifying and recruiting native consultants, co-facilitating workshop sessions with the principal investigator and assisting in e.g. audio or video recording. In some cases, staff members or students were instrumental in digitizing translations, transcriptions or glosses recorded in non-digital formats. 79 native consultants, representing the 59 language varieties participated in the collaborative elicitation workshops (and in some cases in individual elicitation sessions). These individuals offered plenty of linguistic and sociolinguistic insights in interaction with the principal investigator and with other participants, and contributed language data in audio and video recording as well as through writing. To the extent they were able, the consultants made draft (mainly non-IPA) transcriptions, translated portions of text into a language of wider communication, and provided word glossing. Thanks also go to John Peterson (University of Kiel) for invaluable advice and input on feature aggregation and SplitsTree visualizations.

References

Baart, Joan L. G. 1997. *The sounds and tones of Kalam Kohistani: With wordlist and texts* (Studies in Languages of Northern Pakistan 1). Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.

Baart, Joan L. G. 1999. *A sketch of Kalam Kohistani grammar* (Studies in Languages of Northern Pakistan 5). Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.

Baart, Joan L. G. 2014. Tone and stress in North-West Indo-Aryan: A survey. In Johanneke Caspers, Yiya Chen, Willemijn Heeren, Jos Pacilly, Niels O. Schiller & Ellen van Zanten (eds.), *Above and beyond the segments*, 1–13. Amsterdam: John Benjamins.

Bashir, Elena. 1996a. Mosaic of tongues: Quotatives and complementizers in Northwest Indo-Aryan, Burushaski, and Balti. In William L. Hanaway & Wilma Heston (eds.), *Studies in Pakistani popular culture*, 187–286. Lahore: Lok Virsa Pub. House and Sang-e-Meel Publications.

Bashir, Elena. 1996b. The areal position of Khowar: South Asian and other affinities. In Elena Bashir & Israr-ud-Din (eds.), *Proceedings of the second international Hindukush cultural conference* (Hindukush and Karakoram Studies 1), 167–179. Karachi: Oxford University Press.

Bashir, Elena. 2003. Dardic. In George Cardona & Danesh Jain (eds.), *The Indo-Aryan languages*, 818–894. London: Routledge.

Bashir, Elena. 2007. Contact-induced change in Khowar. In Heather Bolton & Saeed Shafqat (eds.), *New perspectives on Pakistan: Visions for the future*, 205–238. Oxford: Oxford University Press.

Bashir, Elena. 2009. Wakhi. In Gernot Windfuhr (ed.), *The Iranian languages*, 825–862. London: Routledge.

Bashir, Elena. 2016. Pre-1947 convergences. In Hans Henrich Hock & Elena Bashir (eds.), *The languages and linguistics of South Asia: A comprehensive guide*, 264–284. Berlin & Boston: De Gruyter Mouton.

Beck, Simone & Daniela Beyer. 2013. A sociolinguistic assessment of the Darwazi speech variety in Afghanistan. *Linguistic Discovery* 11(1). 22–82.

Berger, Hermann. 1998. *Die Burushaski-Sprache von Hunza und Nager 1. Grammatik* (Neuindische Studien 13). Wiesbaden: Harrassowitz.

Bielmeier, Roland. 1985. *Das Märchen vom Prinzen Čobzaṅ: Eine tibetische Erzählung aus Baltistan: Text, Übersetzung, Grammatik und westtibetisch vergleichendes Glossar* (Beiträge zur Tibetischen Erzählforschung 6). Sankt Augustin: VGH Wissenschaftsverlag.

Buddruss, Georg. 1959. *Beiträge zur Kenntnis der Pašai-Dialekte* (Abhandlungen für die Kunde des Morgenlandes XXXIII, 2). Wiesbaden: Steiner.

Burrow, Thomas. 1973. *The Sanskrit language*, 3rd edn. London: Faber and Faber.

Cacopardo, Alberto & Augusto Cacopardo. 2001. *Gates of Peristan: History, religion and society in the Hindu Kush* (Reports and Memoirs 5). Rome: Istituto Italiano per l'Africa e l'Oriente (IsIAO).

Chafe, Wallace L. 1980. *The pear stories: Cognitive, cultural and linguistic aspects of narrative production* (Advances in Discourse Processes, 0896-470X; 3). Norwood, N.J.: Ablex.

Dahl, Östen. 2001. Typological characterization of language families and linguistic areas. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals/Sprachtypologie und sprachliche Universalien/La typologie des langues et les universaux linguistiques. 2. Halbband*, 1456–1470. Berlin & New York: Walter de Gruyter.

Degener, Almuth. 2002. The Nuristani languages. In Nicholas Sims-Williams (ed.), *Indo-Iranian languages and peoples (Proceedings of the British Academy 116)*, 103–117. Oxford: Oxford University Press.

Di Carlo, Pierpaolo. 2011. Two clues of a former Hindu-Kush linguistic area? In Carol Everhard & Elizabeth Mela-Athanasopoulou (eds.), *Selected papers from the international conference on language documentation and tradition – with a special interest in the Kalasha of the Hindu Kush Valleys, Himalayas, 7–9 november 2008*, 101–114. Thessaloniki: School of English, Department of Theoretical & Applied Linguistics, Aristotle University of Thessaloniki.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68(1). 81–138.

Èdel'man, Džoi Iosifovna. 1980. K substratnomu naslediju central'no-aziatskogo jazykovogo sojuza [Towards the substrate heritage of the Central Asian linguistic area]. *Voprosy jazykoznanija* 5. 21–32.

Èdel'man, Džoi Iosifovna. 1983. *The Dardic and Nuristani languages* (Languages of Asia and Africa). Moscow: Nauka.

Èdel'man, Džoi Iosifovna & Leila R. Dodykhudoeva. 2009. The Pamir languages. In Gernot Windfuhr (ed.), *The Iranian languages*, 773–786. London: Routledge.

Enfield, Nick J. & Bernard Comrie (eds.). 2015. *Languages of mainland Southeast Asia, the state of the art*. Berlin & Boston: De Gruyter Mouton.

Heegård, Jan & Henrik Liljegren. 2018. Geomorphic coding in Palula and Kalasha. *Acta Linguistica Hafniensia* 50(2). 129–160.

Heegård Petersen, Jan. 2015. Kalasha texts – With introductory grammar. *Acta Linguistica Hafniensia* 47(1 Suppl). 1–275.

Hill, Nathan W. 2007. Aspirated and unaspirated voiceless consonants in Old Tibetan. *Language and Linguistics* 8(2). 471–493.

Huson, Daniel H. & David J. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267.

International Phonetic Association. 1999. *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Jettmar, Karl. 1975. *Die Religionen des Hindukusch* (Die Religionen Der Menschheit; Bd. 4, 1). Stuttgart: W. Kohlhammer.

Kieffer, Charles M. 2009. Parachi. In Gernot Windfuhr (ed.), *The Iranian languages*, 693–720. London: Routledge.

Koptjevskaja-Tamm, Maria. 2010. Linguistic typology and language contact. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 568–590. Oxford: Oxford University Press.

Koptjevskaja-Tamm, Maria & Henrik Liljegren. 2017. Semantic patterns from an areal perspective. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 204–236. Cambridge: Cambridge University Press.

Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. Circum-Baltic languages: An areal-typological approach. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *The Circum-Baltic languages: Grammar and typology*, 615–750. Amsterdam: John Benjamins.

Lamuwal, Abd-El-Malek & Adam Baker. 2013. Southeastern Pashayi. *Journal of the International Phonetic Association* 43(02). 243–246.

Lehr, Rachel. 2014. *A descriptive grammar of Pashai: The language and speech community of Darrai Nur*. Chicago: University of Chicago PhD Dissertation.

Liljegren, Henrik. 2014. A survey of alignment features in the Greater Hindukush with special references to Indo-Aryan. In Pirkko Suihkonen & Lindsay J. Whaley (eds.), *On diversity and complexity of languages spoken in Europe and North and Central Asia* (Studies in Language Companion Series 164), 133–174. Amsterdam: John Benjamins.

Liljegren, Henrik. 2016. *A grammar of Palula* (Studies in Diversity Linguistics 8). Berlin: Language Science Press.

Liljegren, Henrik. 2019. Gender typology and gender (in)stability in Hindu Kush Indo-Aryan languages. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity I: General issues and specific studies* (Studies in Diversity Linguistics 26), 279–328. Berlin: Language Science Press.

Liljegren, Henrik & Afsar Ali Khan. 2016. Khowar. *Journal of the International Phonetic Association* 47(2). 219–229.

Liljegren, Henrik & Erik Svärd. 2017. Bisyndetic contrast marking in the Hindukush: Additional evidence of a historical contact zone. *Journal of Language Contact* 10(3). 450–484.

Masica, Colin P. 2001. The definition and significance of linguistic areas: Methods, pitfalls, and possibilities (with special reference to the validity of South Asia as a linguistic area). In Peri Bhaskararao (ed.), *The yearbook of South Asian languages and linguistics 2001*, 205–267. London: SAGE.

Mørch, Ida E. 1997. *Retroflexe vokalers oprindelse i kalashamon i historisk og areallingvistisk perspektiv (The origin of retroflex vowels in Kalashamon from a historical and areal-linguistic perspective)*. Copenhagen: University of Copenhagen MA thesis.

Morgenstierne, Georg. 1938. *Indo-Iranian frontier languages Iranian Pamir languages (Yidgha-Munji, Sanglechi-Ishkashmi and Wakhi)* (Instituttet for Sammenlignende Kulturforskning. Serie B, Skrifter, 0332-6217; 35), vol. 2. Oslo: Universitetsforlaget.

Morgenstierne, Georg. 1950. *Notes on Gawar-Bati*. Oslo: Det Norske Videnskaps-Akademi.

Morgenstierne, Georg. 1961. Dardic and Kafir languages. In Peri J. Bearman, Thierry Bianquis, Clifford E. Bosworth, Emeri van Donzel, Wolfhart P. Heinrichs (eds.), *Encyclopedia of Islam*, 2nd edn., vol. 2, Fasc. 25, 138–139. Leiden: Brill.

Morgenstierne, Georg. 1967. *Indo-Iranian frontier languages. Vol. 3, The Pashai language, 1, Grammar (Instituttet for Sammenlignende Kulturforskning. Serie B, Skrifter, 0332-6217; 40:3:1)*. Oslo: Aschehoug.

Morgenstierne, Georg. 1974. Languages of Nuristan and surrounding regions. In Karl Jettmar & Lennart Edelberg (eds.), *Cultures of the Hindukush: Selected papers from the Hindu-Kush cultural conference held at Moesgård 1970* (Beträge zur Südasienforschung 1), vol. 1, 1–10. Wiesbaden: Steiner.

Muysken, Pieter. 2008. Introduction: Conceptual and methodological issues in areal linguistics. In Pieter Muysken (ed.), *From linguistic areas to areal linguistics* (Studies in Language companion series 90), 1–23. Amsterdam: John Benjamins.

Nelson, David N. 1986. *The historical developement of the Nuristani languages*. Minneapolis & St. Paul: University of Minnesota PhD Dissertation.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26(1). 359–384.

Nikolaev, Dmitry & Eitan Grossman. 2018. Areal sound change and the distributional typology of affricate richness in Eurasia. *Studies in Language* 42(3). 562–599.

Payne, John R. 1989. Pamir languages. In Rüdiger Schmitt (ed.), *Compendium linguarum Iranicarum*, 417–444. Wiesbaden: Reichert.

Reichl, Karl. 1983. Syntactic interference in Afghan Uzbek. *Anthropos* 78(3/4). 481–500.

Schmidt, Ruth Laila & Razwal Kohistani. 2008. *A grammar of the Shina language of Indus Kohistan* (Beiträge zur Kenntnis Südasiatischer Sprachen und Literaturen 17). Wiesbaden: Harrassowitz.

Skjaervø, Prods Oktor. 1989. Yidgha and Munji. In Rüdiger Schmitt (ed.), *Compendium linguarum Iranicarum*, 411–416. Wiesbaden: Reichert.

Stassen, Leon. 2013. Zero copula for predicate nominals. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Stilo, Donald L. 2009. Circumpositions as an areal response: The case study of the Iranian zone. *Turkic Languages* 13(1). 3–33.

Tikkanen, Bertil. 1988. On Burushaski and other ancient substrata in northwestern South Asia. *Studia Orientalia* 64. 303–325.

Tikkanen, Bertil. 1999. Archaeological–linguistic correlations in the formation of retroflex typologies and correlating areal features in South Asia. In Roger Blench & Matthew Spriggs (eds.), *Archaeology and language IV: Language change and cultural transformation*, 138–148. London: Routledge.

Tikkanen, Bertil. 2008. Some areal phonological isoglosses in the transit zone between South and Central Asia. In Israr-ud-Din (ed.), *Proceedings of the third international Hindu Kush cultural conference*, 250–262. Karachi: Oxford University Press.

Toporov, Vladimir N. 1970. About the phonological typology of Burushaski. In Jakobson Roman & Shigeo Kawamoto (eds.), *Studies in general and oriental linguistics presented to Shiro Hattori on the occasion of his sixtieth birthday*, 632–647. Tokyo: TEC Corporation for Language and Educational Research.

Toporov, Vladimir N. 1971. Burushaski and Yeniseian languages: Some parallels. In Ivan Poldauf (ed.), *Études de la phonologie, typologie et de la linguistique générale* (Travaux linguistiques de Prague 4), 107–125. Prague: Academia.

Wendtland, Antje. 2009. The position of the Pamir languages within East Iranian. *Orientalia Suecana* 58. 172–188.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. The ASJP database (version 17). http://asjp.clld.org/ (accessed 19 December 2017).

Wilkins, David. 1999. The 1999 demonstrative questionnaire: “this” and “that” in comparative perspective. In David Wilkins (ed.), *Manual for the 1999 field season*, 1–24. Nijmegen: Max Planck Institute for Psycholinguistics.

Zeisler, Bettina. 2005. On the position of Ladakhi and Balti. In John Bray (ed.), *Ladakhi histories: Local and regional perspectives* (Brill’s Tibetan Studies Library 9), 41–64. Leiden: Brill.

Zemp, Marius. 2018. *A grammar of Purik Tibetan* (Brill’s Tibetan Studies Library 21). Leiden: Brill.

Zoller, Claus Peter. 2005. *A grammar and dictionary of Indus Kohistani: volume 1, dictionary* (Trends in Linguistics 21–1). Berlin: Mouton de Gruyter.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jsall-2021-2027).