A TYPOLOGICAL INVESTIGATION OF NEPALESE LANGUAGES

Mark Donohue

The languages of Nepal are established as belonging to four families, with the recent addition of Austroasiatic speakers in the east. This paper moves away from language classification into genealogical families, and examines the classification of the languages of Nepal by examining their morphosyntactic features and applying computational methods.

Keywords: typology, computational analysis, Nepal, comparative morphosyntax.

1. Introduction

Language classification according to descent follows from the application of the comparative method to appropriate data (lexical, or morphological). Typological classification examines languages to determine similarities in structure, without requiring similarities of the sort that lead to judgements of cognacy.

In this paper I examine 59 languages spoken primarily in Nepal, and compare and interpret the results of the clustering analysis performed using Splitstree (Huson and Bryant 2006).

2. The database

The data used is a set of coded features approximately corresponding to the morphosyntactic features present in the World Atlas of Language Structures (WALS; Haspelmath and Dryer 2013). Features were, wherever possible, recoded to present binary features. Thus, for instance, the feature 'Order of Subject and Verb', which has three categorial values in the original WALS coding, was recoded as two features, 'SV' and 'VS', both binary valued. The three values of the original feature can be coded with different combinations of plusses and minuses for the two binary features. After very poorly attested features were discarded (and any not involving morphosyntax), there were 282 recoded features. Of these, 53 were not contrastive for the languages of Nepal; for instance, there are no languages of Nepal for which VS is the basic order of subject and verb, and so this feature was discounted. The remaining 229 features were at least minimally contrastive.

Table 1: Recoding of a *WALS* feature

Original (1 feature)	Recoded (2 features)
Subject precedes verb (SV)	SV (+/–)
Subject follows verb (VS)	VS (+/–)
Both orders with neither dominant	

Examining the database, there were 59 languages primarily spoken in Nepal for which at least 50% of the original set of 282 features were coded, and these languages were the ones examined. The languages involved are from four families, Indo-European,

Dravidian, Tibeto-Burman, and the language isolate Kusunda. The Indo-European languages considered were Nepali, spoken Nepali, Majhi, Darai, Rana Tharu, and Saptariya Tharu. The lone Dravidian language in the sample was Kurukh (Oraon). The Tibeto-Burman languages included varieties from a number of different subgroups: Tibetic (Dolpo, Jirel, Lamjung Yolmo, Lhomi, Mustangi, Nubri (western), Nubri (eastern), Sherpa, Sherpa (Hile), Tsum, and Yohlmo (Helambu), Tamangic (Chantyal, Gurung, Manange, Nar-Phu, Seke, Central Tamang, Western Tamang, Eastern Tamang, Thakali), other Bodic (Dhimal), Newaric (Kathmandu and Dolakha Newar, plus Baram and Thangmi), Western Himalayan (Raji, Raute), the so-called 'Kiranti' languages (Athpare, Bahing, Bantawa, Belhare, Camling, Dumi, Hayu, Jero, Khaling, Koyi, Limbu, Sunwar, Thulung, Wambule, Yakkha, Yamphu), plus Chepang and Bankariya, Dura, Ghale, Kaike, Kham, Kuke, Magar. This is not an exhaustive listing of Nepalese languages, but constitutes those languages which were in the database and which were sufficiently coded to allow for meaningful typological classification. The locations of the different languages are shown in Figure 1, coded according to their position in Figure 1.

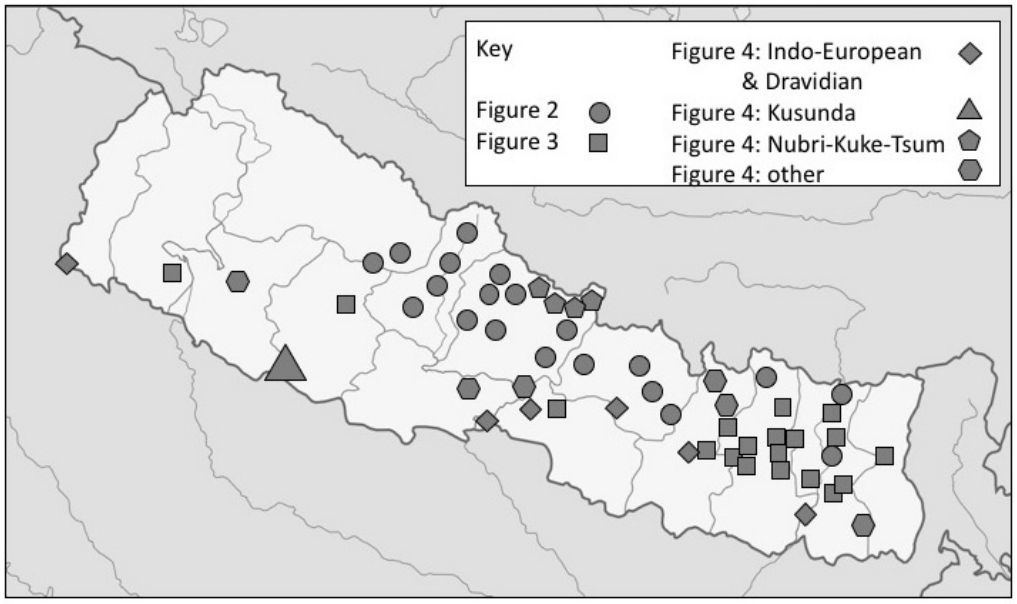

Figure 1: Map showing locations of the languages analysed

3. Methods

It is not possible to compare 229 features between a single pair of languages, let alone 59. Splitstree is a software package designed to take large amounts of data from a large number of sample points, and present the results of various clustering analyses in the form of a two-dimensional phylogram. Since the input data represents typological features, the output will show the degree of similarity between any one language and the others in the dataset. It is not possible for a language to appear as *not*structurally related

to the other languages in the database, but we can assess the different levels of similarity. A neighbornet representation allows for a non-categorial display of relationships, and since typological traits are known to be subject to diffusion, this presents a more realistic way to evaluate similarities between different languages and be able to interpret the results. I will first present the results of the analysis of the 59 languages in the database that are sufficiently well coded, and the focus on the emergent low-level clusters.

4. Results

Figure 2 presents the neighbournet analysis of all 59 languages. It is clear that there are two sharply defined parts of the graph, one clearly separated from the rest at the top, and one less clearly defined onthe right. They both have about the same number of languages, one third of the dataset each, and the remaining languages form a non-clustering third part of the data.

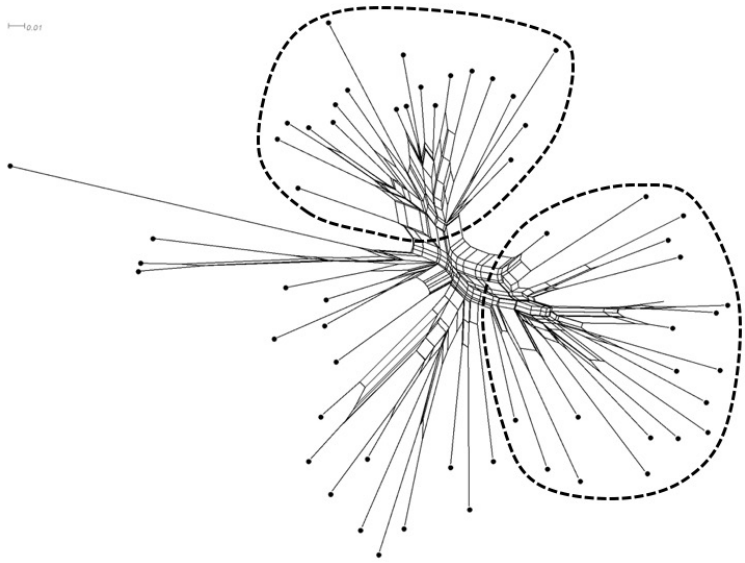

Figure 2: Neighbournet analysis of 59 languages

Taking the top third of the graph in detail, we can see how these languages relate to each other. The languages in this branch of the graph are all Tibeto-Burman, and are predominantly from two genealogical groups: Tibetan, and Tamangic. In addition there are two Newaric languages, Baram and Dura, and the Kaike and Ghale languages, which have a complicated language history. The Tibetic languages are typologically similar, with Dolpo, Mustang, Yolmo (Helambu and Lamjung), and Sherpa from Hile all forming a typological subgroup together. Included in this group are the non-Tibetic languages of Kaike and Nar-Phu; both of these languages are in extensive contact with Tibetic languages, and have clearly been affected by them structurally.

At the bottom of Figure 3 we find most of the Tamangic languages: three varieties of Tamang, Gurung, Seke, Thakali, Manange, and peripherally Chantyal, but also the

Tibetic languages of Lhomi (more central) Sherpa. Compared to the other Tibetic languages in Figure 3, Lhomi and (central) Sherpa have been much more strongly influenced by the languages of the Nepalese hills, away from the Tibetan plateau and a linguistic ecology that reinforces the Tibetic typology.

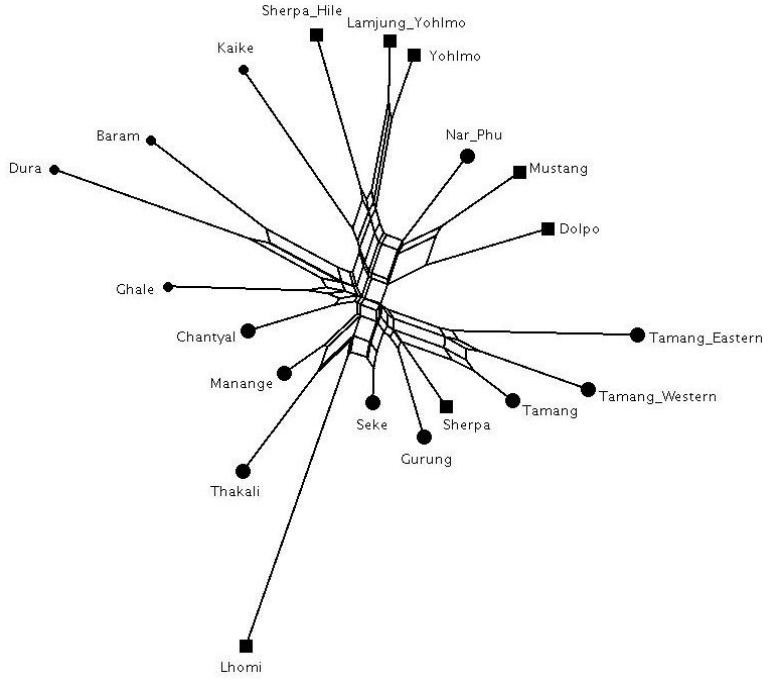

Figure 3: Neighbournet analysis of 20 languages at the top of Figure 2

The less clearly defined group on the bottom right of Figure 2 is shown in detail in Figure 4. These 20 languages are mostly from the Kiranti group (though see Ebert 2003 for qualification of this label), with all the languages included showing a shared head-marking typology that is not found in most other languages of Nepal. The presence of Chepang and Kham in this typological group, and the failure of Limbu to separate out from the various Rai languages, emphasises the point that typological analysis does not replicate linguistic lineages (Donohue and Musgrave 2007). We further note that almost none of the smaller clusters within this group reflect accepted subgroups of Kiranti.

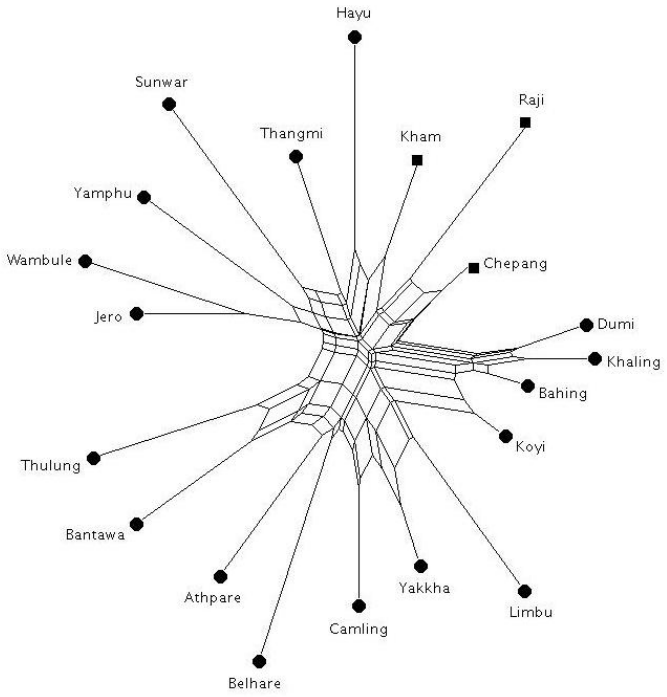

Figure 4: Neighbournet analysis of 20 languages at the right of Figure 2

The remaining 19 languages from Figure 2 do not form a coherent group, but have been treated together for the purposes of exemplification. The languages to the far left of Figure 2 are Kuke, Kusunda, and the two Nubri varieties. These languages appear at the bottom of Figure 5, and clearly stand out from the other languages of Nepal. As can be seen in Figure 5, Kusunda is quite typologically distinct from the other languages in the study.There is a clearly defined cluster consisting of the Indo-European languages, on the left of Figure 5, with Kurux tenuously linked to this group as part of a pan-South Asian typological group distinct from the structural norms of the mostly Tibeto-Burman languages of the Himalayas. Trailing towards the Kiranti group to the right in Figure 2 we have the disparate Tibeto-Burman languages Dhimal, Raute, Bankariya and Thangmi, which do not form a group against any other languages, but (as with the rest of the non-Indo-European languages in Figure 5 apart from Kuke and Nubri) are a set of 'left-over' languages that are not part of the major typological subgroups that have emerged from the analysis. At the bottom right of Figure 5 we can see the typological position of the languages of Upper Gorkha: Nubri varieties (Samagaun and Prok) cluster together, closely joined by Kuke, and tenuously connected with Tsum.

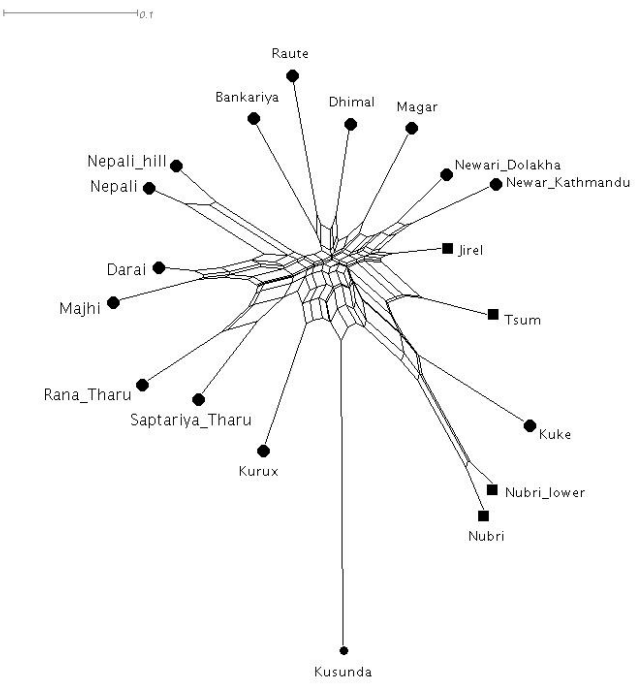

Figure 5. Neighbournet analysis of 19 languages at the bottom-left of Figure 2

5. Interpretation

As mentioned above, certain discrepancies between accepted linguistic phylogenies and the clusters arising from computational typological analysis. The appearance of Sherpa and Lhomi embedded within the (sub-)cluster that is made up of central hills Tamangic languages shows that extensive contact with non-Tibetic languages can affect a languages (morphosyntactic) typology to the point that it no longer behaves like its close relatives; similarly, extensive contact with Tibetan languages has led the language of Nar and Phu to more closely resemble these languages than its Tamangic relatives. At the same time, the appearance of both Tibetic and Tamangic languages in a single cluster, most distinct from the other languages in the sample, indicates shared language history in the form of genealogical links, or areal diffusion. There is no surprise in the suggestion that Tibetic and Tamangic languages form a subgroup (e.g., Thurgood 2003), but it is all but certain that Baram does not constitute part of a subgroup including the Tamangic and Tibetic languages (the position of Ghale and Kaike are more contentious, and the affiliations of Dura have not yet been elucidated). The geographic proximity of Ghale, Kaike, Baram and Dura to the other languages in this part of the graph makes it very likely that extensive language contact has affected the structure of these four languages to the point that they now behave, typologically, more like the Tamangic (Ghale) or Tibetic (Kaike) languages than any other group in Nepal.

In Figure 4 we saw the clustering confusion that arises in a geographically and typologically tightly knit group of languages from eastern Nepal, with clear evidence of multiple layers of typological sharing. The presence of Chepang, Kham and Raji in this group, despite their lack of genealogical or geographic affinity with the Kiranti languages, and in the absence of these three languages themselves forming a genealogical or geographic grouping, suggests that the typological profile typified by the Kiranti languages was previously more widespread than it is today, and that the eastern exemplars of this stuctural profile represent relics of an earlier time during which head-marking and prefixation was more widespread in the region that is now Nepal.

The 'remnant' languages which have been shown in detail in Figure 5 offer much for interpretation and speculation. The first point of note is Kusunda, typologically distinct from the other languages in Figure 5 and those in Figure 2. Kusunda is an isolate, but that alone does not require that it be typologically distinct from its neighbours (witness the similarity of Nihali and Burushaski to South Asian typological norms). Kusunda, nevertheless, presents a very different morphosyntactic profile that is very different from the other languages of Nepal (e.g., Donohue and Gautam 2013, Donohue, Gautam and Pokharel 2014, Gautam and Donohue 2014, Watters et. al. 2005, and other work). Within the Indo-European cluster on the left of Figure 5 we can see that 'Hill Nepali', the variety of Nepali used by non-Chetri-Bahun speakers, is still firmly in the Indo-European cluster, but is leaning towards the Tibeto-Burman languages (in Figure 5, it leans to the left of the Indo-European cluster, towards Newari). Here, too, we can see the effects of language contact in this quantified methodology we can detect the substratal effects on the morphosyntactic typology of languages.

The position of the Nubri varieties, Kuke, and Tsum is very interesting. Nubri and Tsum are lexically Tibetic languages; nonetheless, their morphosyntactic profile does not fit the the Tibetic languages in Figure 3, but represents a continuum of variation towards the non-ergative profile of Kusunda.

6. Discussion

Further work will involve the comparison of phonological features, as well as morphosyntactic ones, to determine whether or not, and to what extent, these different datasets result in different clusters amongst the languages compared (see Donohue 2014). It has been suggested (McConvell 2002, 2008) that nominal morphosyntax and verbal morphosyntax show different patterns of diffusion, and reflect different kinds of contact scenarios, so a separation of nominal and verbal traits would both allow for a test on this hypothesis, and potentially greater insights into recent and ancient language contact scenarios in Nepal.

Further, it is clear that the addition of extra languages would allow for more perspective on the bounds of areality. These languages would be selected from areas close to Nepal and as representatives of the families involved (eg., more plateau Tibetan languages to offer a genealogical perspective, more southerly Indo-European languagesaway from the Himalayas, and also Tibeto-Burman langauges from both west of and east of Nepal.

References

Dhakal, Dubi Nanda, Mark Donohue, Bhojraj Gautam, and Naijing Liu. 2016. Diagnosing a contact history for Tsum. *Nepalese Linguistics* 31. 14-20.

Donohue, Mark. 2013. Who inherits what, when? contact, substrates and superimposition zones. In Balthasar Bickel, David Peterson, Lenore Grenoble and Alan Timberlake, eds., *Language Typology and Historical Contingency; in honor of Johanna Nichols*. 219-239. Typological Studies in Language 104. Amsterdam: John Benjamins.

Donohue, Mark. 2014. Studying contact without detailed studies of the languages involved: a non-philological approach to language contact. In Kayla Carpenter, Oana David, Florian Lionnet, Christine Sheil, Tammy Stark and Vivian Wauters, eds, *Proceedings of the 38th Annual meeting of the Berkeley Linguistic Society, (2012) (Approaches to Language Contact)*. 92-120. Berkeley: Berkeley Linguistic Society.

Donohue, Mark, Bhoj Raj Gautam and Madhav Prasad Pokharel. 2014. Negation and nominalisation in Kusunda. *Language* 90(3). 737-745.

Donohue, Mark, and Simon Musgrave. 2007. Typology and the linguistic macro-history of island Melanesia. *Oceanic Linguistics* 46(2). 348-387.

Donohue, Mark, and Bhoj Raj Gautam. 2013. Evidence and stance in Kusunda. *Nepalese Linguistics* 28. 38-47.

Donohue, Mark, Simon Musgrave, Bronwen Whiting and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. *Language* 87(2). 369-383.

Dryer, Matthew S., and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Ebert, Karen H. 2003. Kiranti Languages: An Overview. In Graham Thurgood and Randy LaPolla, eds, *The Sino-Tibetan Languages*. 505-517. London: Routledge.

Gautam, Bhoj Raj and Mark Donohue. 2014. Deixis in Kusunda. *Nepalese Linguistics* 29. 152-157.

Huson, Daniel H., and David Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23(2). 254-267.

McConvell, Patrick. 2002. 'Mix-im-up' speech and emergent mixed languages in indigenous Australia. *Proceedings of SALSA 2001. Texas Linguistic Forum* 44(1-2). 328-349.

McConvell, Patrick. 2008. Mixed Languages as Outcomes of Code-Switching: Recent Examples from Australia and Their Implications. Journal of Language Contact 2(1). 187-212.

Thurgood, Graham. 2003. A Subgrouping of the Sino-Tibetan Languages: The Interaction between Language Contact, Change, and Inheritance. In Graham Thurgood and Randy LaPolla, (eds.) *The Sino-Tibetan Languages*. 3-21. London: Routledge.

Watters, David E., with Yogendra P. Yadava, Madhav P. Pokharel and Balaram Prasain. 2006. *Notes on Kusunda grammar*. Himalayan Linguistics Archive 3. 1-182.