Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics

Lars Borin♣, Anju Saxena♡, Taraka Rama♣, Bernard Comrie♢♠

♣University of Gothenburg
Gothenburg, Sweden
lars.borin@svenska.gu.se
taraka.rama.kasicheyanula@gu.se

♡Uppsala University
Uppsala, Sweden
anju.saxena@lingfil.uu.se

♢Max Planck Institute
for Evolutionary Anthropology
Leipzig, Germany
♠University of California
Santa Barbara, USA
comrie@eva.mpg.de

Abstract

Like many other research fields, linguistics is entering the age of big data. We are now at a point where it is possible to see how new research questions can be formulated – and old research questions addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small, carefully selected samples. For example, South Asia is often mentioned in the literature as a classic example of a linguistic area, but there is no systematic, empirical study substantiating this claim. Examination of genealogical and areal relationships among South Asian languages requires a large-scale quantitative and qualitative comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. We present some preliminary results of our large-scale investigation of the genealogical and areal relationships among the languages of this region, based on the linguistic descriptions available in the 19 tomes of Grierson's monumental *Linguistic Survey of India* (1903–1927), which is currently being digitized with the aim of turning the linguistic information in the LSI into a digital language resource suitable for a broad array of linguistic investigations.

Keywords: South Asian languages, genetic linguistics, areal linguistics

1. Introduction

Like many other research fields, linguistics is entering the age of big data. The modern digital world and the mass digitization of historical documents together provide unprecedented opportunities to linguistics and other disciplines relying on text and speech as primary research data. However, this development comes with considerable methodological challenges. We are now at a point where it is possible to see how new research questions can be formulated – and old research questions addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small, carefully selected samples, but where a methodology has not yet established itself, and where serious studies have hardly been conducted at all.

For example, comprehensive, large-scale quantitative and qualitative studies are essential in order to get a deeper understanding of areal linguistics. South Asia[1] is often mentioned in the literature as a classic example of a linguistic area. There is, however, no systematic, empirical study of South Asian languages to substantiate this claim. In order to critically evaluate South Asia as a linguistic area, a systematic examination of a set of linguistic features in a wide range of South Asian languages is essential.

South Asian languages belong to four major language families: Indo-European (>Indo-Aryan), Dravidian, Austroasiatic (>Mon-Khmer and Munda), and Sino-Tibetan (>Tibeto-Burman). There are also some small families (e.g., in the Andaman Islands), some language isolates (e.g., Burushaski and Nihali), and some unclassified languages.

Throughout history multilingualism has been the norm in the area. There are signs of language contact between Vedic Sanskrit and Dravidian languages in the Rig Veda, the oldest text found in India. It has been claimed that this long-lasting contact situation has made the languages of this region more similar in some respects to each other than they are to their genealogically related languages spoken outside this region, and that consequently South Asia should be seen as a linguistic area (e.g., Emeneau 1956; Masica 1976; Kachru et al. 2008, and others).

However, systematic investigations of this claim have been few and somewhat spotty, mostly relying on data from a few major Indo-Aryan and Dravidian languages (see Ebert 2006 for a critique). The approach of Subbarao (2008) is representative: Linguistic features (most of them from Emeneau 1956) are illustrated with single – 'cherry-picked' – linguistic examples, and different languages are used to illustrate different linguistic features. This is understable at one level: One would like to include as many languages as possible in a study, and doing the work manually puts severe restrictions as to how many languages and/or features one can handle.

In order to critically evaluate the notion of South Asia as a linguistic area, we need to know the spread and extent of a linguistic feature across space and language families. Further, the internal sub-grouping of all the South Asian language families remains unclear. E.g., Asher (2008) problematizes the current internal subgrouping of the Indo-

[1] Although South Asia is defined variously in the literature, in linguistic works this area is usually considered to comprise the seven countries Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka.

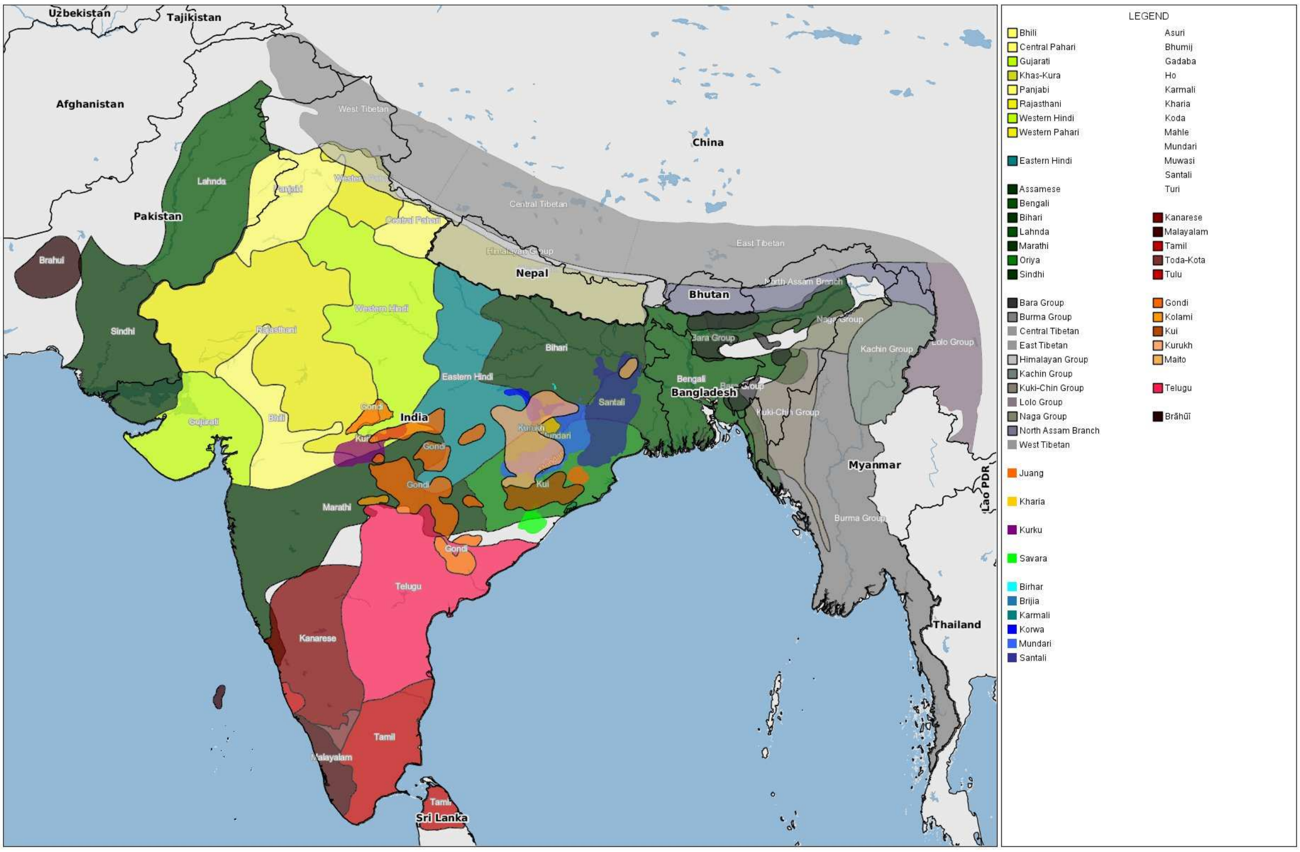


Figure 1: Map of four major South Asian language families (from `http://llmap.org`)

Aryan language family, as the proposed subgroups correlate highly with their geographical distribution.[2] The focus in works on the internal relationships is on one family at a time, e.g., Turner (1964), Bloch (1954), Cardona and Jain (2003) on Indo-Aryan; Burrow and Emeneau (1984), Krishnamurti (2003) on Dravidian; Matisoff (2003) and Thurgood and LaPolla (2003) on Sino-Tibetan. However, given the claims about South Asia as a linguistic area, it would be prudent to always have an eye open for contact influences from other families, since these might vitiate aspects of a purely family-internal investigation.

The map in figure 1 shows the geographical extent of the four major South Asian language families (Austroasiatic, Dravidian, Indo-Aryan, and Sino-Tibetan), including overlap between languages belonging to different families. Three prominent cases of such overlap are Gondi (Dravidian)–Marathi (Indo-Aryan), Brahui (Dravidian)–Sindhi (Indo-Aryan), and Santali (Munda)–Bengali (Indo-Aryan).

2. Towards a language resource from Grierson's LSI

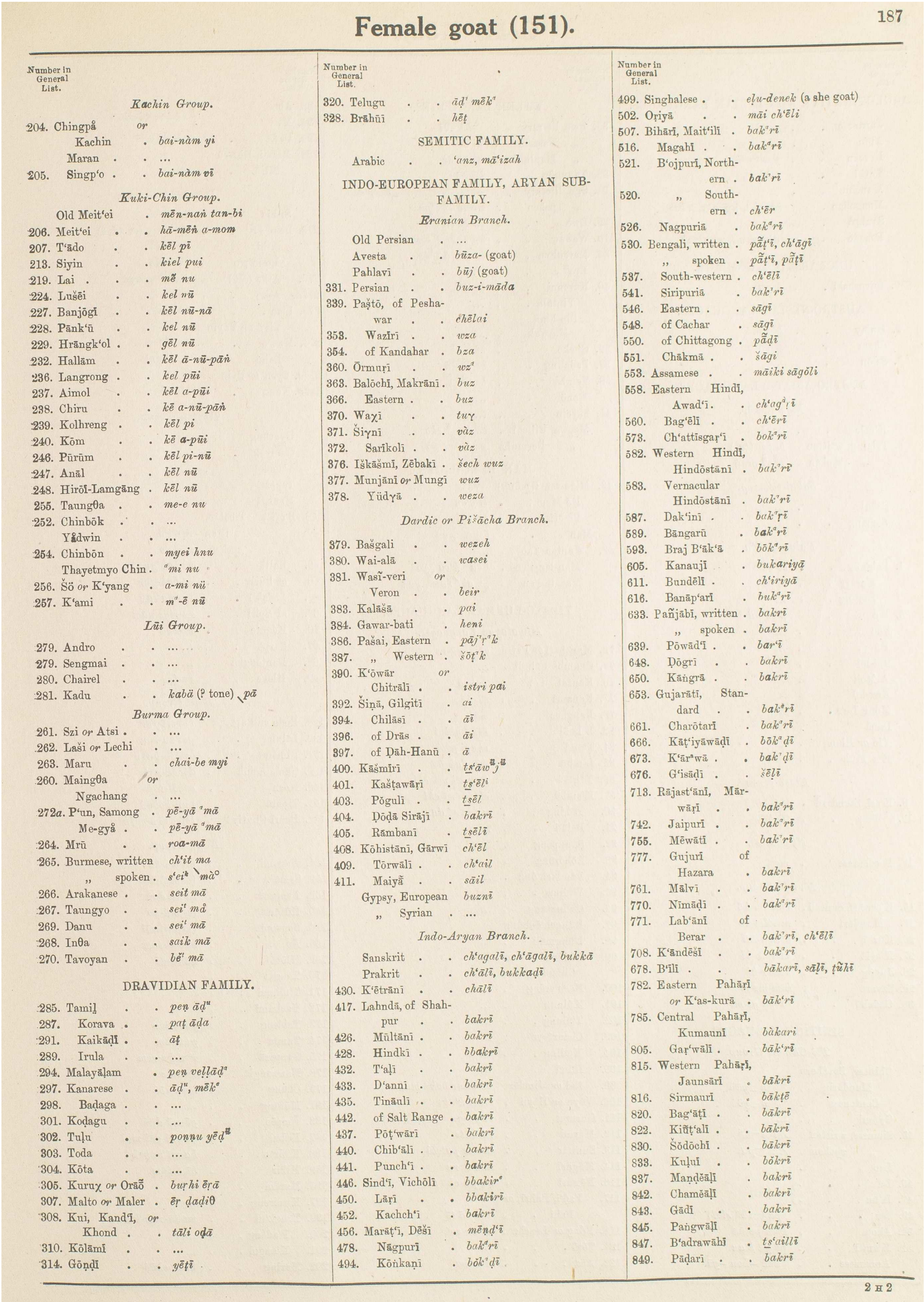
Examination of genealogical and areal relationships among South Asian languages requires a large-scale comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools.

In this paper we will present some preliminary results of our large-scale investigation of the genealogical and areal relationships among the languages of this region, based on the linguistic material available in Grierson's *Linguistic Survey of India* (LSI; Grierson 1903–1927), which is currently being digitized with the aim of turning the linguistic information in the LSI into a digital language resource, a database suitable for a broad array of linguistic investigations, which will be made freely available under an open-content license.[3]

The LSI still remains the most complete single source on South Asian languages. Its 19 tomes (9500 pages) cover 723 linguistic varieties representing major language families and some unclassified languages, of almost the whole of nineteenth-century British-controlled India (modern India, Pakistan, Bangladesh, and parts of Burma). For each

[2] Correlation with geographical distribution is not *inherently* a problem, of course. If a number of speech communities have migrated minimally since the split-up of their ancestor language, then one would expect geography and genealogy to correlate. But this then makes it hazardous to attempt to distinguish between effects of geography and genealogy on the basis of language-internal evidence alone.

[3] See, e.g., the IDS wordlists (Borin et al., 2013) produced in our ongoing project, available under a CC-BY license from `http://spraakbanken.gu.se/eng/research/digital-areal-linguistics/word-lists`.



187

Female goat (151).

Number in General List.		
	Kachin Group.	
204.	Chingpå *or* Kachin	*bai-nàm yi*
	Maran	...
205.	Singp'o	*bai-nàm vī*
	Kuki-Chin Group.	
	Old Meit'ei	*mēn-naṅ tan-bi*
206.	Meit'ei	*hā-mēṅ a-mom*
207.	T'ādo	*kēl pī*
213.	Siyin	*kiel pui*
219.	Lai	*mē̃ nu*
224.	Lušēi	*kel nū*
227.	Banjōgī	*kēl nū-nā*
228.	Pānk'ū	*kel nū*
229.	Hrāngk'ol	*gēl nū*
232.	Hallām	*kēl ā-nū-pāṅ*
236.	Langrong	*kel pūi*
237.	Aimol	*kēl a-pūi*
238.	Chiru	*kē a-nū-pāṅ*
239.	Kolhreng	*kēl pi*
240.	Kōm	*kē a-pūi*
246.	Pūrūm	*kēl pi-nū*
247.	Anāl	*kēl nū*
248.	Hirōī-Lamgāng	*kēl nū*
255.	Taungθa	*me-e nu*
252.	Chinbōk	...
	Yådwin	...
254.	Chinbōn	*myei hnu*
	Thayetmyo Chin	^{a}*mi nu*
256.	Šö *or* K'yang	*a-mi nü*
257.	K'ami	*m^{a}-ē nū*
	Lūi Group.	
279.	Andro	...
279.	Sengmai	...
280.	Chairel	...
281.	Kadu	*kabä* (? tone) \pā
	Burma Group.	
261.	Szi *or* Atsi	...
262.	Laši *or* Lechi	...
263.	Maru	*chai-be myi*
260.	Maingθa *or* Ngachang	...
272*a*.	P'un, Samong	*pē-yā ^{a}mā*
	Me-gyå	*pē-yā ^{a}mā*
264.	Mrū	*roa-mā*
265.	Burmese, written	*ch'it ma*
	,, spoken	*s'ei^{k} \\mà°*
266.	Arakanese	*seit mā*
267.	Taungyo	*sei^{t} må*
269.	Danu	*sei^{t} mā*
268.	Inθa	*saik mā*
270.	Tavoyan	*bĕ^{t} mā*
	DRAVIDIAN FAMILY.	
285.	Tamiḻ	*peṇ āḍ^{u}*
287.	Korava	*paṭ āḍa*
291.	Kaikāḍī	*āṭ*
289.	Irula	...
294.	Malayāḷam	*peṇ veḷḷāḍ^{a}*
297.	Kanarese	*āḍ^{u}, mēk^{e}*
298.	Baḍaga	...
301.	Koḍagu	...
302.	Tuḷu	*poṇṇu yēḍ^{ŭ}*
303.	Toda	...
304.	Kōta	...
305.	Kuruχ *or* Orāõ	*buṛhi ēṛā*
307.	Malto *or* Maler	*ēṛ ḍaḍiθ*
308.	Kui, Kand'ī, *or* Khond	*tāli oḍā*
310.	Kōlāmī	...
314.	Gōṇḍī	*yēṭī*

Number in General List.		
320.	Telugu	*āḍ^{i} mēk^{a}*
328.	Brāhūī	*hēṭ*
	SEMITIC FAMILY.	
	Arabic	*'anz, mā'izah*
	INDO-EUROPEAN FAMILY, ARYAN SUB-FAMILY.	
	Eranian Branch.	
	Old Persian	...
	Avesta	*būza-* (goat)
	Pahlavī	*būj* (goat)
331.	Persian	*buz-i-māda*
339.	Paṣ̌tō, of Peshawar	*ćhēlai*
353.	Wazīrī	*wza*
354.	of Kandahar	*bza*
360.	Ōrmuṛī	*wz^{a}*
363.	Balōchī, Makrānī	*buz*
366.	Eastern	*buz*
370.	Waχī	*tuγ*
371.	Šiγnī	*vàz*
372.	Sarīkolī	*vàz*
376.	Iškāšmī, Zēbakī	*šech wuz*
377.	Munjānī *or* Mungī	*wuz*
378.	Yüdγā	*weza*
	Dardic or Piśācha Branch.	
379.	Bašgali	*wezeh*
380.	Wai-alā	*wasei*
381.	Wasī-veri *or* Veron	*beir*
383.	Kalāšā	*pai*
384.	Gawar-bati	*heni*
386.	Pašai, Eastern	*pāj^{i}ṛ^{a}k*
387.	,, Western	*šōṭ^{a}k*
390.	K'ōwār *or* Chitrālī	*istri pai*
392.	Šiṇā, Gilgitī	*ai*
394.	Chilāsī	*āī*
396.	of Drās	*āi*
397.	of Ḍāh-Hanū	*ā*
400.	Kāšmīrī	*t̲s'āw^{ŭ}j^{ŭ}*
401.	Kaštawāṛī	*t̲s'ēl^{i}*
403.	Pōgulī	*t̲sēl*
404.	Ḍōḍā Sirājī	*bakrī*
405.	Rāmbanī	*t̲sēlī*
408.	Kōhistānī, Gārwī	*ch'ēl*
409.	Tōrwālī	*ch'ail*
411.	Maiyā̃	*sāil*
	Gypsy, European	*buznī*
	,, Syrian	...
	Indo-Aryan Branch.	
	Sanskrit	*ch'agalī, ch'āgalī, bukkā*
	Prakrit	*ch'ālī, bukkaḍī*
430.	K'ētrānī	*chālī*
417.	Lahndā, of Shahpur	*bakrī*
426.	Mūltānī	*bakrī*
428.	Hindkī	*bbakrī*
432.	T'aḷī	*bakrī*
433.	D'annī	*bakrī*
435.	Tināulī	*bakrī*
442.	of Salt Range	*bakrī*
437.	Pōṭ'wārī	*bakrī*
440.	Chib'ālī	*bakrī*
441.	Punch'ī	*bakrī*
446.	Sind'ī, Vichōlī	*bbakir^{e}*
450.	Lāṛī	*bbakirī*
452.	Kachch'ī	*bakrī*
456.	Marāṭ'ī, Dēšī	*mēṇḍ'ī*
478.	Nāgpurī	*bak^{a}rī*
494.	Kōṅkaṇī	*bŏk^{a}ḍī*

Number in General List.		
499.	Singhalese	*eḷu-denek* (a she goat)
502.	Oṛiyā	*māi ch'ēli*
507.	Bihārī, Mait'ilī	*bak^{a}rī*
516.	Magahī	*bak^{a}rī*
521.	B'ojpurī, Northern	*bak^{i}rī*
520.	,, Southern	*ch'ēr*
526.	Nagpuriā	*bak^{a}rī*
530.	Bengali, written	*pā̃ṭ'ī, ch'āgī*
	,, spoken	*pā̃ṭ'ī, pā̃ṭī*
537.	South-western	*ch'ēlī*
541.	Siripuriā	*bak^{i}rī*
546.	Eastern	*sāgī*
548.	of Cachar	*sāgī*
550.	of Chittagong	*pā̃ḍī*
551.	Chākmā	*šāgi*
553.	Assamese	*māiki sāgŏli*
558.	Eastern Hindī, Awad'ī	*ch'ag^{a}ṛī*
560.	Bag'ēlī	*ch'ērī*
573.	Ch'attīsgaṛ'ī	*bok^{a}rī*
582.	Western Hindī, Hindōstānī	*bak^{i}rī*
583.	Vernacular Hindōstānī	*bak^{i}rī*
587.	Dak'inī	*bak^{a}ṛī*
589.	Bāngarū	*bak^{a}rī*
593.	Braj B'āk'ā	*bōk^{a}rī*
605.	Kanaujī	*bukariyā̤*
611.	Bundēlī	*ch'iriyā*
616.	Banāp'arī	*buk^{a}rī*
633.	Pañjābī, written	*bakrī*
	,, spoken	*bakrī*
639.	Pōwād'ī	*bar'ī*
648.	Ḍōgrī	*bakrī*
650.	Kāṅgrā	*bakrī*
653.	Gujarātī, Standard	*bak^{a}rī*
661.	Charōtarī	*bak^{a}rī*
666.	Kāṭ'iyāwāḍī	*bōk^{a}ḍī*
673.	K'ār^{a}wā	*bak^{i}ḍī*
676.	G'isāḍī	*šēḷī*
713.	Rājast'ānī, Mārwāṛī	*bak^{a}rī*
742.	Jaipurī	*bak^{a}rī*
755.	Mēwātī	*bak^{i}rī*
777.	Gujurī of Hazara	*bakrī*
761.	Mālvī	*bak^{i}rī*
770.	Nīmāḍī	*bak^{a}rī*
771.	Lab'ānī of Berar	*bak^{i}rī, ch'ēlī*
708.	K'āndēšī	*bak^{a}rī*
678.	B'īlī	*bākarī, sāḷī, ṭū̃hī*
782.	Eastern Pahāṛī *or* K'as-kurā	*bāk'rī*
785.	Central Pahāṛī, Kumaunī	*bàkari*
805.	Gaṛ'wālī	*bāk'rī*
815.	Western Pahāṛī, Jaunsārī	*bākrī*
816.	Sirmaurī	*bākṭē*
820.	Bag'āṭī	*bākrī*
822.	Kiū̃ṭ'alī	*bākrī*
830.	Šŏdōchī	*bākrī*
833.	Kuḷuī	*bŏkrī*
837.	Maṇḍēāḷī	*bakrī*
842.	Chamēāḷī	*bakrī*
843.	Gādī	*bakrī*
845.	Paṅgwāḷī	*bakrī*
847.	B'adrawāhī	*t̲s'aillī*
849.	Pāḍarī	*bakrī*

2 H 2

Figure 2: The LSI comparative vocabulary.

major variety it provides (1) a grammatical sketch (including a description of the sound system); (2) a core word list; and (3) texts (including a translation of the Parable of the Prodigal Son). The core word lists which accompany the language descriptions are collected in a separate volume (Volume 1, Part 2: *Comparative vocabulary*; see figure 2).

Each list has a total of 168 entries (concepts). The concepts in the comparative vocabulary cover a broad spectrum consisting of body parts, domestic animals, personal pronouns, numerals, and astronomical objects.

There is some overlap with other concept lists used in language classification: First, 38 of the concepts are also found in the shorter (100-item) version of the so-called *Swadesh lists*, core vocabulary lists originally devised by the American linguist Morris Swadesh (1950; 1952; 1955) specifically for the purpose of inferring genealogical relationships among languages.

Further, 76 of the items are found in an extended Swadesh list used by us in earlier genealogical investigations of Tibeto-Burman languages of the Indian Himalayas (e.g. Saxena 2011; Saxena and Borin 2011; Saxena and Borin 2013). Similarly, 34 LSI vocabulary items are present in the Leipzig-Jakarta list, a 100-item list of word senses claimed to be highly resistant to borrowing (Haspelmath and Tadmor, 2009).

Thus, the LSI comparative vocabulary clearly has one part that can be used in investigating genetic connections among the languages, but also another part – at least half of the entries – which we hypothesize could be used to find areal influences.

Family	# varieties
Austro-Asiatic	12
Dravidian	19
Indo-Aryan	96
Sino-Tibetan	141

Table 1: Major South Asian family languages in the LSI comparative vocabulary

3. Some preliminary experiments

In this paper, we focus on the data extracted from the comparative vocabulary. All in all, the LSI offers core vocabulary lists for more than 250 language varieties from the four main South Asian language families (see table 1).

The wide range of languages and the number and type of concepts present in this language resource allow us address the issue of genealogical vs. areal factors using computational methods (Wichmann, 2008). These computational methods take a set of vocabulary lists as input, and yield a distance matrix between the vocabulary lists as output. The distance matrix may subsequently be used as an input to a phylogenetic program to infer a classification tree for the set of languages.

The distance matrix is computed through the application of a variant of Levenshtein distance (Levenshtein, 1966), LDND (Levenshtein Distance Normalized Double; Wichmann et al. 2010). The computation of LDND between a pair of word lists is carried out as follows:

LDN is computed as the sum of the Levenshtein distance between the words occupying the same meaning slot, normalized by length. Similarity between phoneme inventories and chance similarity might cause a pair of not-so related languages to show up as related languages. This is compensated for by computing the length-normalized Levenshtein distance between all the pairs of words occupying different meaning slots and summing the different word-pair distances.

The summed Levenshtein distance between the words occupying the same meaning slots is divided by the sum of Levenshtein distances between different meaning slots. The intuition behind this idea is that if two languages are shown to be similar (small distance) due to accidental chance similarity then the denominator would also be small and the ratio would be high.

If the languages are not related and also share no accidental chance similarity, then the distance as computed in the numerator would be unaffected by the denominator. If the languages are related then the distance as computed in the numerator is small anyway, whereas the denominator would be large since the languages are similar due to genetic relationship and not from chance similarity. Hence, the final ratio would be smaller than the original distance given in the numerator.

The matrix is then given to a Neighbor-Joining program (Saitou and Nei, 1987) to yield an unrooted tree. For our data, this cross-family tree turns out to group the languages into distinct clusters corresponding to the recognized language families.

Subsequently, we evaluate the overall accuracy of classification in each language family through a comparison of a family's distance matrix with its gold-standard family tree, extracted from the latest edition of the *Ethnologue* (Lewis et al., 2013).[4] An example of such a tree, for the Dravidian language family, is shown in figure 3.

One way to compare the distance matrix with the family tree is to compute the correlation between pairwise language distances and the pairwise branch lengths (as read off from the tree). However, it is not obvious how to best compute the distances from the family tree.

The most straightforward method would be the *raw branch length*, or simply the number of nodes encountered in the shortest path lying from a language A to a language B. The left-most plot in figure 4 shows the agreement between LDND distance and pair-wise raw branch length distance for the Dravidian language family. The fit is not particularly good. There could be more than one reason for this, of course, but the naïve raw branch length method is certainly a strong suspect, given the standard assumption that

[4] The Ethnologue is not above reproach as a gold standard, but at present there is hardly a better comprehensive source of language family information covering such a broad range of languages.

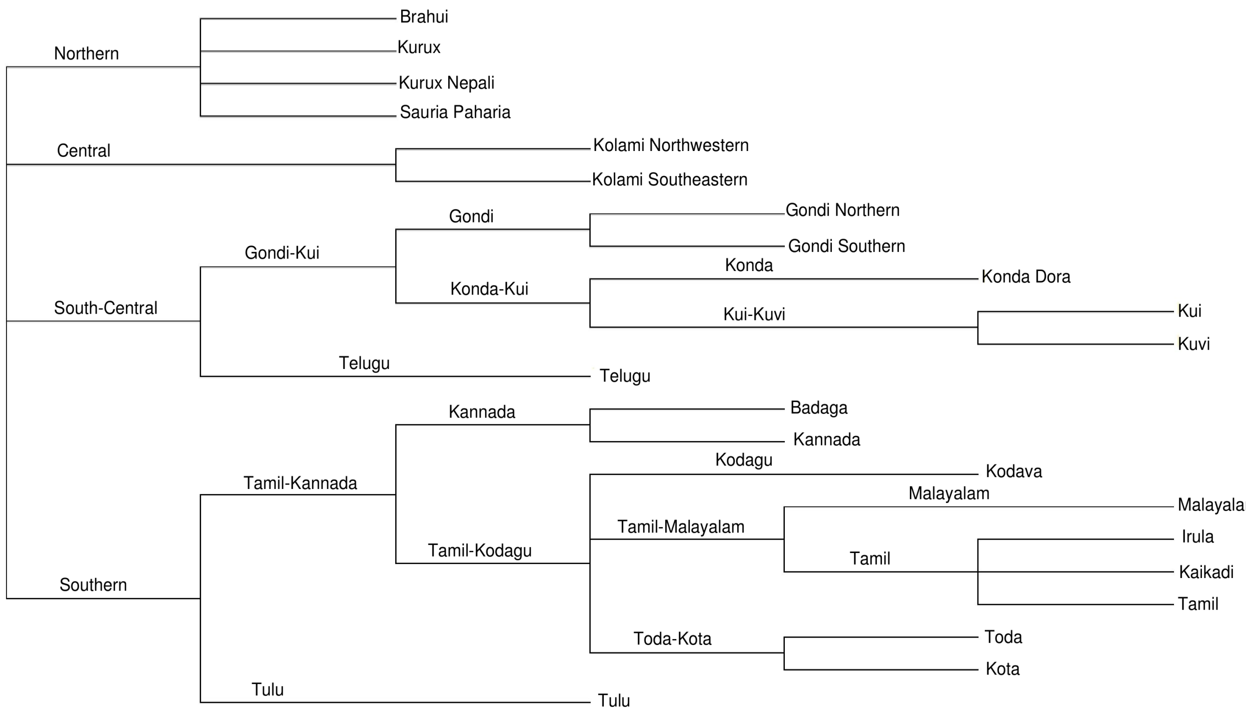


Figure 3: The Dravidian family tree according to *Ethnologue*

observed language change – and consequently the distance between related language varieties – should be a function of time depth.

If we then assume that the horizontal dimension in figure 3 reflects time depth, all terminal nodes in the family tree should lie at the same distance from the root node, since they all represent present-day languages with an equally long history of descent from the proto-language.

Typical language family trees are 'unbalanced', with a different number of intermediate nodes on different branches, as the Dravidian tree in figure 3, where the Northern Dravidian language Brahui is much closer to the root than Southern Dravidian Malayalam.

We propose to balance the Ethnologue trees in the following way: Each node is weighted according to its depth from the root node, with the maximal depth (the terminal node furthest away from the root) always set to be 1. For instance, the Northern Dravidian node would get a weight of 0.5 (1/2) as opposed to the Southern Dravidian node which is assigned a weight of 0.17 (1/6). The rightmost plot in figure 4 shows the agreement when the branch lengths are computed from the balanced Dravidian family tree.

The agreement between the distance matrix and the branch length matrix can be computed using Kendall's τ, a rank-based correlation measure highly suited for this purpose, since it takes ties in ranks into account.[5] Kendall's τ computes the level of agreement between the language-pair LDND distances and the tree distance derived from the Ethnologue tree.

Given the LDND distances $ldnd_i, ldnd_j$ and the gold standard distances gsd_i, gsd_j for language pairs i, j where, $1 <= i, j <= n(n-1)/2$ and n is the number of languages. Kendall's τ counts a i, j pair as concordant if $ldnd_i < ldnd_j$ and $gsd_i < gsd_j$ and discordant, if $ldnd_i > ldnd_j$ and $gsd_i < gsd_j$ or if $ldnd_i < ldnd_j$ and $gsd_i > gsd_j$. A pair i, j is counted as tie when $gsd_i = gsd_j$ or $ldnd_i = ldnd_j$.

As can be seen from figure 4, a number of language pairs share the same tree distance. Finally, τ is defined as the ratio of the difference between the number of concordant pairs and discordant pairs to the square root of the product of the number of non-tied pairs in both $ldnd$ and gsd. The value of τ lies between -1 and $+1$ where, -1 suggests a perfect disagreement whereas $+1$ suggests a perfect agreement between the quantities under comparison.

There is a significant ($p < 0.001$) improvement of τ from the use of raw branch length to balanced tree branch length. The statistical significance of the scores is calculated using a Mantel test (Mantel, 1967), a permutation test which computes the significance of a test statistic by permuting the rows of a matrix and recomputing the τ score.

The Mantel test counts the number of times the observed correlation is greater than or lesser than the correlation computed between permuted matrices. The idea behind this test is that if the observed correlation is by chance then there is equally likely chance that the correlation between the permuted matrices is greater than or lesser than the observed correlation.

If the permuted correlations are always lesser than or greater than the observed correlation then the test result can be interpreted as highly significant. Usually, a permu-

[5] Because of the tree topology, there will normally be more than one language pair with the same distance between them.

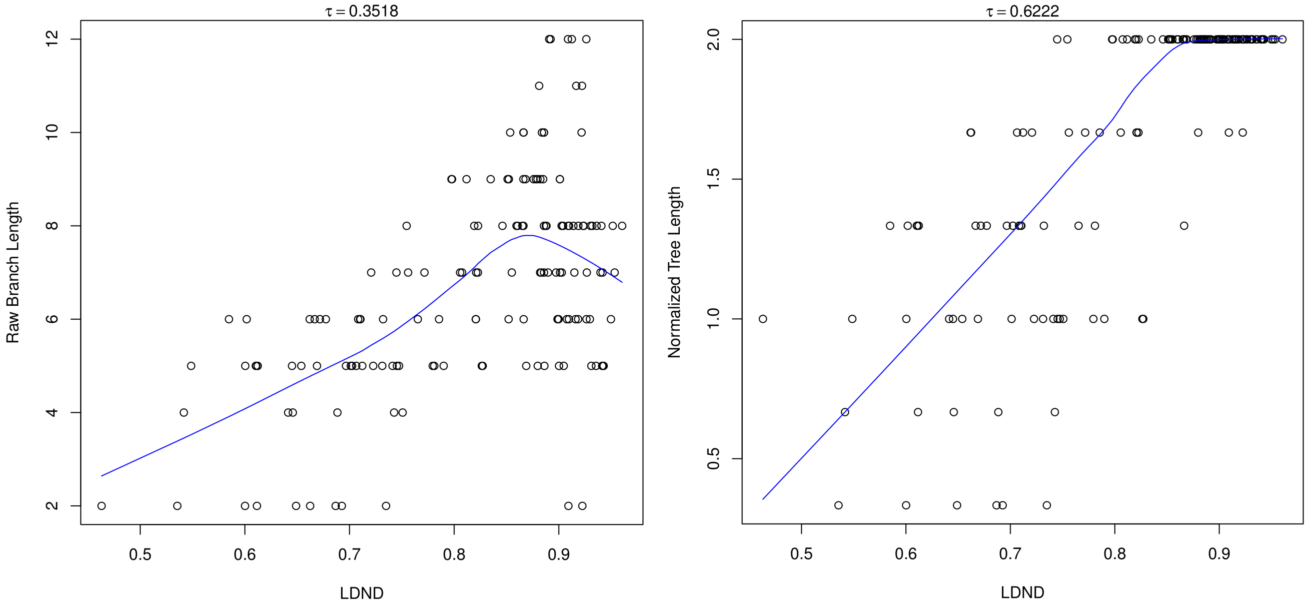


Figure 4: Distance matrix fit to Dravidian raw tree (left) and balanced tree (right) shown by locally fitted regression lines.

tation of 1000 times is sufficient to estimate the reliability of the correlation. In our experiments, we permuted the matrices 1000 times and noticed that the chance that the observed correlation is accidental is less than or equal to 0.001. The mantel test experiments have been run using the *vegan* package implemented in R (R Core Team, 2012).

Family	τ raw	τ balanced
Austroasiatic	0.5187	0.6219
Dravidian	0.3518	0.6222
Indo-Aryan	0.4032	0.4576
Sino-Tibetan	0.2745	0.3577

Table 2: Kendall's τ for the four major South Asian language families

Another way of computing the agreement between the gold-standard tree and the LDND distances is offered by a modified version of Goodman-Kruskal's γ (Goodman and Kruskal, 1954). This score compares language triplets with regard to whether both measures show the same distance relations among the languages in a triplet or not (ignoring ties). γ lies in the range $[-1, 1]$ where a score of -1 indicates total disagreement and a score of $+1$ indicates perfect agreement.

The computation of γ is illustrated with an example from the Dravidian language family. The program for computing γ searches through all the language triplets and checks if there is a 'closest pair' in the triplet, i.e. a pairing of the languages in the triplet where the members of the pair are closer to each other than each of them is to the third language. For example, in the language triplet, Kannada–Tulu–Telugu: Kannada and Tulu are closer to each other than Telugu in the family tree (see figure 3).

Since there are three pairwise LDND distances in this language triplet, the program makes two comparisons: if the LDND distance between Kananda–Tulu is lesser than the LDND distance between Kannada–Telugu as well as Tulu–Telugu. An agreement is counted as a concordant comparison and a disagreement is counted as a discordant comparison. The ratio of the difference between the number of concordant comparisons and the number of discordant comparisons to the ratio of the sum of concordant comparisons and the number of discordant comparisons in all the triplets yields γ. It has to be noted that ties are ignored in the computation. Brahui–Telugu–Tulu is an example of a tie triplet since the pair-wise distances between all the languages is the same.

We computed γ for the four major South Asian language families and found that LDND agrees with the balanced branch length score better than the raw branch length score (see table 3).

Family	γ raw	γ balanced
Austroasiatic	0.6379	0.7766
Dravidian	0.4259	0.748
Indo-Aryan	0.4819	0.6514
Sino-Tibetan	0.4189	0.5885

Table 3: Goodman-Kruskal's γ for the four major South Asian language families

4. Discussion, conclusions and outlook

The experiments described above have shown how LDND distance calculations on the LSI comparative vocabulary recover both inter-family and intra-family genealogical relations for the four major South Asian language families. However, no indications of areal phenomena could be seen using this method on the LSI comparative vocabulary. The phylogenetic trees built from the distance matrix cluster related languages together, whereas no cross-family areal clusters emerge.

The correlation between LDND distances and balanced family tree distances was high for both measures used, whereas with a strong areal component, a lower correlation would have been expected.

There could be several conceivable reasons for this, e.g.:

(1) There is in fact no areal effect to be recovered from the data;
(2) the comparison method chosen (LDND) is not suitable for this problem;
(3) contrary to our expectations (see section 2), the LSI comparative vocabulary is the 'wrong' vocabulary for uncovering language contact; or
(4) we need to look at other parts of the language than vocabulary in order to establish areal connections.

Given the amount and quality of the argumentation advanced in support of the hypothesis of South Asia as a linguistic area, it will take much stronger counterevidence than the results presented in this paper to even begin to falsify this hypothesis. Hence, (1) is not a reasonable assumption at this point.

As for (2), on the one hand, in a broad comparison of different string similarity measures with respect to their effectiveness for genealogical language comparison (Rama and Borin, forthcoming), LDND came out on top for the task of calculating the internal genealogical classification of a language family, but is was one of the least effective measures out of those evaluated for distinguishing unrelated from related languages. Clearly, more research is needed in this area.

In other work, we have made some experiments using a more linguistically informed semi-automatic vocabulary comparison showing very encouraging results (Saxena, 2011; Saxena and Borin, 2011; Saxena and Borin, 2013). This may certainly be an avenue worth exploring, and future research will show if this methodology will scale up sufficiently to deal with the LSI comparative vocabulary.

Unfortunately, the LSI cannot help us with (3), should this turn out to be the case. There is in fact some evidence in the literature (e.g. Gumperz and Wilson 1971), that in intensive contact situations such as those which give rise to linguistic areas, grammar may be affected more than vocabulary. Hence, with respect to (4), we are in the process of extracting the information on the various grammatical features found in the LSI grammar sketches into a rich typological database, which will hopefully provide us with a firmer basis for investigating areal and micro-areal phenomena in South Asia.

5. References

Asher, Ronald E. (2008). Language in historical context. In Kachru, Braj B., Kachru, Yamuna, and Sridhar, S. N., editors, *Language in South Asia*, pages 31–46. Cambridge University Press, Cambridge.

Bloch, Jules. (1954). *The grammatical structure of Dravidian languages*. Deccan College, Pune. Authorized translation from the original French by Ramkrishan Ganesh Harshé.

Borin, Lars, Comrie, Bernard, and Saxena, Anju. (2013). The Intercontinental Dictionary Series – a rich and principled database for language comparison. In Borin, Lars and Saxena, Anju, editors, *Approaches to Measuring Linguistic Differences*, pages 285–302. De Gruyter Mouton, Berlin.

Burrow, Thomas and Emeneau, Murray B. (1984). *A Dravidian etymological dictionary*. Clarendon Press, Oxford, 2nd edition.

Cardona, George and Jain, Dhanesh. (2003). *The Indo-Aryan languages*. Routledge, London.

Ebert, Karen. (2006). South Asia as a linguistic area. In Brown, Keith, editor, *Encyclopedia of languages and linguistics*. Elsevier, Oxford, 2nd edition edition.

Emeneau, Murray B. (1956). India as a linguistic area. *Language*, 32:3–16.

Goodman, Leo A. and Kruskal, William H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Grierson, George A. (1903–1927). *A Linguistic Survey of India*, volume I-XI. Government of India, Central Publication Branch, Calcutta.

Gumperz, John J. and Wilson, Robert. (1971). Convergence and creolization: A case from the Indo-Aryan/Dravidian border in India. In Hymes, Dell, editor, *Pidginization and Creolization of Languages*, pages 151–167. Cambridge University Press, Cambridge.

Haspelmath, Martin and Tadmor, Uri, editors. (2009). *Loanwords in the world's languages: A comparative handbook*. De Gruyter Mouton.

Kachru, Braj B., Kachru, Yamuna, and Sridhar, S. N., editors. (2008). *Language in South Asia*. Cambridge University Press, Cambridge.

Krishnamurti, Bhadiraju. (2003). *The Dravidian languages*. Cambridge University Press, Cambridge.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

Lewis, M. Paul, Simons, Gary F., and Fennig, Charles D., editors. (2013). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Mantel, Nathan. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Masica, Colin P. (1976). *Defining a linguistic area: South Asia*. University of Chicago Press, Chicago.

Matisoff, James A. (2003). *Handbook of Proto-Tibeto-Burman. System and philosophy of Sino-Tibetan reconstruction*. University of California Press, Berkeley.

R Core Team, (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rama, Taraka and Borin, Lars. (forthcoming). Comparative evaluation of string similarity measures for automatic language classification. In Mačutek, Ján and

Mikros, George K., editors, *Sequences in Language and Text*. De Gruyter Mouton.

Saitou, Naruya and Nei, Masatoshi. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Saxena, Anju. (2011). Towards empirical classification of Kinnauri varieties. In Austin, Peter K., Bond, Oliver, Nathan, David, and Marten, Lutz, editors, *Proceedings of Conference on Language Documentation & Linguistic Theory 3*, London. SOAS.

Saxena, Anju and Borin, Lars. (2011). Dialect classification in the Himalayas: A computational approach. In *NODALIDA 2011 Conference Proceedings*, pages 307–310, Riga. NEALT.

Saxena, Anju and Borin, Lars. (2013). Carving Tibeto-Kanauri by its joints: Using basic vocabulary lists for genetic grouping of languages. In Borin, Lars and Saxena, Anju, editors, *Approaches to measuring linguistic differences*, number 265 in Trends in Linguistics. Studies and Monographs, pages 175–198. De Gruyter Mouton, Berlin.

Subbarao, Karumuri Y. (2008). Typological characteristics of South Asian languages. In Kachru, Braj B., Kachru, Yamuna, and Sridhar, S. N., editors, *Language in South Asia*, pages 49–78. Cambridge University Press, Cambridge.

Swadesh, Morris. (1950). Salish internal relationships. *International Journal of American Linguistics*, 16:157–167.

Swadesh, Morris. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Swadesh, Morris. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Thurgood, Graham and LaPolla, Randy, editors. (2003). *The Sino-Tibetan languages*. Routledge, London.

Turner, Sir Ralph L. (1964). *A comparative dictionary of the Indo-Aryan languages*. Oxford University Press, Oxford.

Wichmann, Søren. (2008). The emerging field of language dynamics. *Language and Linguistics Compass*, 2(3):442–455.

Wichmann, Søren, Holman, Eric W., Bakker, Dik, and Brown, Cecil H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389:3632–3639.