Dialect Classification in the Himalayas: a Computational Approach

Anju Saxena
Department of Linguistics and Philology
Uppsala University, Sweden
anju.saxena@lingfil.uu.se

Lars Borin
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
lars.borin@svenska.gu.se

Abstract

Linguistic fieldwork data – in the form of basic vocabulary lists – for nine closely related language varieties are compared using an automatic procedure with manual feedback, whose major advantage is its complete consistency. The results of the vocabulary comparison turn out to be in accord with other linguistic features, making this methodology a promising addition to the toolbox of genetic lingusitics.

1 Introduction

The aim of the work presented here is to examine genetic relationships among nine Tibeto-Burman varieties spoken in the Kinnaur region in India, using a semi-automatic computational approach. The focus in this presentation is on lexical items, although grammatical features are also taken into account in our work.

2 Background: Kinnauri varieties and the language data used

The Tibeto-Burman varieties to be discussed here are collectively referred to as Kinnauri and are spoken[1] in the Kinnaur region in the Himachal Pradesh state in India. They belong to the West-Himalayish sub-branch of the Tibeto-Burman language family, which in turn forms one of the two primary subdivisions of the Sino-Tibetan language family. There is brief mention of some Kinnauri varieties in some older works (e.g., Gerard 1842; Cunningham 1844). However, to date there has not been any systematic, comparative linguistic study of the Kinnauri varieties, and consequently no systematic basis for examining how the Tibeto-Burman varieties spoken in Kinnaur relate to one another.

The fieldwork to collect the data used in this investigation was conducted in the following villages in Kinnaur: Nichar (Ni), Sangla (Sa), Chitkul (Ch), Kalpa (Ka), Kuno (Ku), Labrang (La), Poo (Po), Ropa (Ro) and Nako (Na). See figure 1. The main motivation for selecting these villages was to include data from as diverse geographical regions as possible. The data comprise (i) a basic vocabulary list (a revised Swadesh list; Swadesh 1955) for all sites (242 senses); (ii) an extended IDS list for Sangla and Nako (1884 senses);[2] and (iii) selected grammatical constructions.

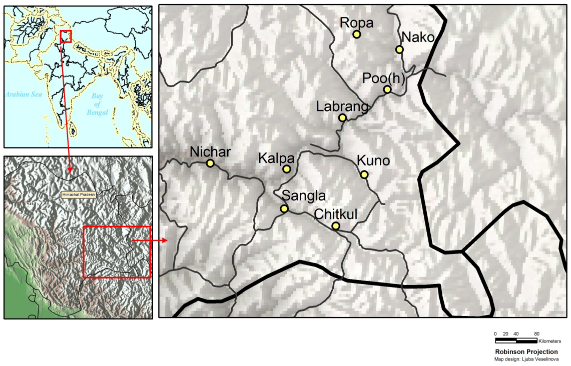

Figure 1: Villages in Kinnaur where data collection was conducted

3 Procedure for word list comparison

The procedure which we have used for comparing the word lists here is similar to recent work in dialectometry (e.g., Nerbonne and Heeringa 2009) and lexicostatistics (e.g., Holman et al. 2008) in relying on a completely automatic comparison of the items in the word lists. However, it differs from most of this work – a notable exception being the work reported on by McMahon et al. (2007) – in its usage of rules tailored to the particular linguistic configuration under investigation, rather than a general method for string comparison. In this respect, it falls somewhere in between traditional glottochronology – where expert statements are required about the cognacy of items – and these modern approaches – which rely entirely on surface form for determining identity of items – although closer to the latter than the former.

[1] None of them have a conventional written form.

[2] See <http://lingweb.eva.mpg.de/ids>. For practical reasons, the IDS list could be collected only for two varieties, and we chose two varieties spoken at the extreme ends of the main river valley running through Kinnaur.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 307–310

	Sa	Ni	Ka	Ro	Ch	La	Po	Ku	Na
new/183*	[1] ɲug	[1] ɲug	[2] ɲuk	[3] ɲʊkʰ:	[4] nʊɪ	[4] nʊɪ	[5] sɔma	[5] sɔma	[5] soma
red/172	[1] ʃʊɪg	[1] ʃʊig	[1] ʃʊig	[1/2] ʃʊig; ʃʊik	[3] məɪ	[3] məi	[4] marpo	[4] marbo	[4] marʊo
small/32	[1/2] dzikts; gaʈo	[2/3] gaʈo; dzɪk	[4] dzɪgits	[4] dzɪgɪts	[5] ats	[6] tsɪgɪts	[7] cʊn:	[7] cʊn:	[7] cun
warm/180	[1] bok	[1/1] bok:; bɔk:	[1] bɔk:	[1] bɔk:	[2/3] tatʰra; ta	[4] kocʰra	[5/6] ɖɔnmo; ʈanmo	[6] ʈɔnmo	[7] ʈonmo
white/175	[1] ʈʰog	[1] ʈʰog	[1] ʈʰog	[1] ʈʰog	[2] tʃaɪn	[3] tʃai	[4] karʊo	[4] karbo	[4] karʊo
yellow/174	[1] pɪg	[1] pig	[2] pik	[2/1] pik; pig	[3] lei	[3] lei	[4] sekʰa	[5] sɛrbo	[5] serʊo

(a) Comparison and assignment to equivalence classes of some adjectives

	Ni	Ka	Ro	Ch	La	Po	Ku	Na
Sa	13/19 (68%)	9/19 (47%)	11/19 (57%)	1/18 (5%)	1/18 (5%)	1/17 (5%)	0/19 (0%)	1/19 (5%)
Ni		12/19 (63%)	10/19 (52%)	1/18 (5%)	1/18 (5%)	0/17 (0%)	0/19 (0%)	0/19 (0%)
Ka			11/19 (57%)	1/18 (5%)	1/18 (5%)	0/17 (0%)	0/19 (0%)	0/19 (0%)
Ro				1/18 (5%)	1/18 (5%)	1/17 (5%)	0/19 (0%)	1/19 (5%)
Ch					8/18 (44%)	0/17 (0%)	0/18 (0%)	1/18 (5%)
La						0/17 (0%)	1/18 (5%)	1/18 (5%)
Po							10/17 (58%)	9/17 (52%)
Ku								11/19 (57%)

(b) Comparison of all adjectives

	Ni	Ka	Ro	Ch	La	Po	Ku	Na
Sa	67/94 (71%)	69/94 (73%)	51/94 (54%)	44/94 (46%)	28/93 (30%)	10/93 (10%)	13/95 (13%)	11/94 (11%)
Ni		61/93 (65%)	42/93 (45%)	36/93 (38%)	25/92 (27%)	9/92 (9%)	9/94 (9%)	8/93 (8%)
Ka			51/93 (54%)	41/93 (44%)	27/92 (29%)	10/92 (10%)	11/94 (11%)	9/93 (9%)
Ro				37/93 (39%)	37/92 (40%)	14/92 (15%)	20/94 (21%)	16/93 (17%)
Ch					29/92 (31%)	8/92 (8%)	11/94 (11%)	10/93 (10%)
La						16/91 (17%)	20/93 (21%)	17/92 (18%)
Po							48/93 (51%)	57/92 (61%)
Ku								53/94 (56%)

(c) Comparison of all nouns

	Ni	Ka	Ro	Ch	La	Po	Ku	Na
Sa	107/157 (68%)	106/155 (68%)	85/161 (52%)	60/158 (37%)	36/156 (23%)	13/156 (8%)	16/159 (10%)	14/162 (8%)
Ni		102/153 (66%)	74/156 (47%)	49/154 (31%)	33/152 (21%)	10/151 (6%)	11/155 (7%)	9/156 (5%)
Ka			86/154 (55%)	54/151 (35%)	34/149 (22%)	11/148 (7%)	13/152 (8%)	10/154 (6%)
Ro				53/157 (33%)	45/155 (29%)	16/154 (10%)	22/158 (13%)	18/160 (11%)
Ch					44/155 (28%)	10/154 (6%)	13/157 (8%)	13/157 (8%)
La						20/152 (13%)	24/156 (15%)	21/155 (13%)
Po							76/154 (49%)	87/155 (56%)
Ku								84/158 (53%)

(d) Comparison of all words in the word list

Table 1: Some results of the comparisons

The main methodological advantage of our approach is its consistency, and not as claimed for the work just referred to, that it should be language-independent. Instead, in our case we try to apply a principle sometimes formulated in computational linguistics as “Don’t guess if you know” (Tapanainen and Voutilainen, 1994, 47), which leads us to include language-specific knowledge in the form of correspondence rules among dialects.

The following proceedure was used in this investigation, developed in collaboration between a computational linguist (Borin) and the linguist who collected the language data (Saxena):

- After the data collection and initial processing of the data,
- a list of observations of relationships among varieties was made by the linguist.
- This list formed the basis for developing a set of principles for comparing the linguistic correspondences in these Kinnauri varieties. These were formulated by the linguist and computational linguist together and their purpose was to determine which segmental differences to disregard for the purpose of considering items in different varietes as the same.
- The principles were encoded by the computational linguist as context-sensitive phonological segment equivalence rules in a small computer program for comparing items fully automatically in order to achieve consistency.
- The equivalence rules were revised after inspection of the result, and the program run again on the data. This process went through a few iterations until the linguist was satisfied with the result.

The results are indicative and sometimes subject to revision, but interesting. They come in the form of two kinds of tables:

- tables of individual lexical items, where items considered the same in different varieties get the same numerical index (table 1a);
- summary tables, where similarities among all lexical items of a particular grammatical or semantic category (nouns, kinship terms, etc.) are shown as ratios and percentages (see tables 1b–1d).

4 Results

The findings of this survey will be illustrated here by focusing on the following lexical sets: adjectives, nouns, numerals and all words. In table 1a, all Swadesh list items are further identified by their Swadesh list number added to the end of the English word and separated from the word by a slash: *small/32*.

Data on adjectives are shown in tables 1a and 1b. Yellow indicates Kinnauri varieties which show 50% or more similarity, blue indicates 10% or less similarity. Sangla, Nichar, Kalpa and Ropa share a higher degree of similarity with one another. Similarly, there is a higher degree of similarity between Poo, Kuno and Nako. But there is very little similarity between the varieties of the Sangla group and the varieties of the Nako group.

Table 1c gives corresponding figures for nouns.

The numerals 1–10 in the Kinnauri varieties are cognates to a very large extent – consistent with the Tibeto-Burman numeral forms noted by Hodson (1913). For the numerals 1, 4, 7, 8 and 10 these varieties use two distinct cognate forms: Poo, Kuno and Nako use the same forms as noted by Hodson (1913) for Central Tibetan, while Nichar, Sangla, Kalpa, Ropa, Chitkul and Labrang use another set of forms.[3] These data, which lack of space prevents us from showing, thus support the pattern emerging from the comparisons that we show here.

5 Summary and discussion

See table 1d. This comparison suggests that

1. the Kinnauri varieties on the two extremes of this table form two separate sub-groups, referred to here as the “Sangla group” at the left end and the “Nako group” at the right end.
2. The core members of the Sangla group are Sangla, Nichar and Kalpa, with Ropa as a more peripheral member. The core members of the Nako group are Poo, Kuno and Nako. They show a high degree of mutual similarity (mostly more than 50%).
3. These tables also display a consistent sharp distinction between the Sangla group and the Nako group, where the degree of similarity between the two groups is less than 10% in most cases.
4. Concerning the status of the remaining two varieties, Chitkul and Labrang:
   a. The degree of similarity between Chitkul and Labrang is neither very high nor very low. It is 28%
   b. Concerning their relationship to the two groups, Chitkul – much more than Labrang – shows a relatively higher degree of similarity with the Sangla group (31–37%) than with the Nako group (6–8%).

[3] A similar subgrouping pattern emerges also concerning the formation of higher numerals in the Kinnauri varieties.

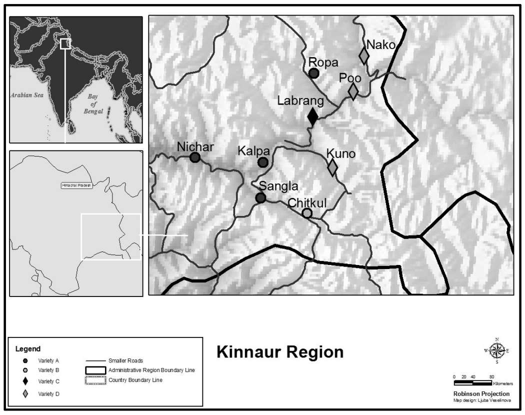

Figure 2: Dialect groups according to our study

c. The status of Labrang is interesting. It shows the highest affinity with Ropa and Chitkul (28–29%) – even though not very high. Labrang does not show much similarity with either group

The systematic comparison of these linguistic features has revealed how the various Kinnauri varieties are similar or dissimilar to one another, thus providing the linguistic basis for examining the relationship among them. The results show that the investigated varieties can be classified into three (or possibly four) groups, where Sangla, Nichar, Ropa and Kalpa form one group; Poo, Kuno and Nako form another group; Chitkul and Labrang fall somewhere in between, being (separately) more to one or the other group concerning some linguistic features, but distinct with regard to other linguistic features. See figure 2.

6 Conclusions and future work

Due to restrictions of space, many details concerning the data collection process, language consultants, geography and demography of Kinnaur, language contact, other investigated linguistic features, etc., have been omitted here. These will be described in future publications from our project.

The automatic vocabulary comparison has yielded good results which are in accord with other linguistic evidence for the genetic linguistic subgrouping of the investigated Kinnauri varieties. A clear methodological advantage is the complete consistency of the comparison. The method will be developed further and its relationship to other similar work investigated, e.g., the work at Groningen on dialectometry (Nerbonne and Heeringa, 2009) as well as work at MPI Leipzig and elsewhere on the theory and methodology of automated large-scale lexicostatistics (Holman et al., 2008; Ringe, 1999; Ringe et al., 2002; Wichmann and Grant, 2010).

Acknowledgements

The research reported on here was supported by the Swedish Research Council (the project *Digital areal linguistics* 2010–2012, VR dnr 2009-1448).

References

Lieutenant J.D Cunningham. 1844. Notes on Moorcroft's travels in Ladakh, and on Gerard's account of Kunáwar, including a general description of the latter district. *Journal of the Asiatic Society of Bengal*, XIII:172–253.

Alexander Gerard. 1842. A vocabulary of the Kunawar languages. *Journal of the Asiatic Society of Bengal (Calcutta*, 11:478–551.

T.C. Hodson. 1913. Note on the numeral systems of the Tibeto-Burman dialects. *Journal of the Royal Asiatic Society of Great Britain and Ireland*, pages 315–336, Apr.

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica*, 42(2):331–354.

April McMahon, Paul Heggarty, Robert McMahon, and Warren Maguire. 2007. The sound patterns of Englishes: Representing phonetic similarity. *English Language and Linguistics*, 11(1):113–142.

John Nerbonne and Wilbert Heeringa. 2009. Measuring dialect differences. In Jürgen Erich Schmidt and Peter Auer, editors, *Language and space: Theories and methods*, pages 550–567. Mouton De Gruyter, Berlin.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Don Ringe. 1999. How hard is it to match CVC-roots? *Transactions of the Philological Society*, 97(2):213–244.

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

Pasi Tapanainen and Atro Voutilainen. 1994. Tagging accurately – don't guess if you know. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 47–52, Stuttgart. ACL.

Søren Wichmann and Anthony P. Grant, editors. 2010. *Quantitative approaches to linguistic diversity. Commemorating the centenary of the birth of Morris Swadesh*. Benjamins. Special issue of *Diachronica* 27:2.