Bandhu Collection Project: Working Paper 2

Aspect markers as an indicator of narrative style in Nepali

Andrew Hardie and Ram Lohani
Department of Linguistics and English Language
Lancaster University

Introduction

This project's first working paper outlined how, supported by a grant from the British Academy,[1] the Bandhu Collection of early-1970s Nepali spoken text was re-digitised in a number of forms (original orthographic form, transliterated Devanagari form, and part-of-speech tagged form). In this working paper, we exemplify one of the types of research for which the Collection may be employed.

In light of evidence from analysis of the Nepali National Corpus, which will be outlined below, there is some reason to believe that *narrativity* (or *narrative content*) is a parameter impacting on grammatical variation across Nepali text types – although, as noted below, this parameter may subsume a number of other dimensions of variation that have been identified in studies on English text-types.

The nature of the Bandhu Collection as a corpus consisting largely of prototypical oral narratives (anecdotes, folk stories, and so on) makes it a highly useful point of contrast to the different written genres of the NNC to throw into relief the grammatical features of narrativity. In this working paper, we will concentrate particularly on the frequency of markers of progressive aspect. However, we will also look briefly at part-of-speech frequency as a more general grammatical parameter of variation.

Part-of-speech frequencies across genres in the Core Sample of the Nepali National Corpus

The NNC has been part-of-speech (POS) tagged using the Nelralec Tagset.[2] This is a group of around 100 word-level morphosyntactic categories. Using this annotated data, it is possible to identify which categories are characteristic of any given division of the corpus in comparison to the rest of the corpus. This is done using a *keyness* analysis based on statistical significance. Any POS tag whose frequency is (relatively) greater in a given subcorpus than in the remainder of the corpus, to a degree that is highly statistically significant,[3] is considered a *key tag* for that subcorpus. The use of POS category frequency as a parameter of grammatical variation in texts is well-

[1] Grant reference: British Academy Small Research Grant SG-42148.
[2] See Hardie et al. (2005, forthcoming)
[3] With a p-value less than 0.01. The log likelihood statistic was used (see Dunning 1993).

established in the literature: see for instance Biber (1988), Mair et al. (2002), and for an overview see Hardie (2007). The specific key POS tags methodology applied here has previously been utilised by Rayson et al. (2002).

The actual divisions on which this analysis was based are built into the sampling frame of the Brown Corpus (see Francis and Kučera 1979) and similar corpora such as LOB and FLOB (see, respectively, Johansson et al. 1978; Hundt et al. 1998). The Core Sample (CS) of the NNC was built on this sampling frame, which specifies fifteen narrowly-defined text categories within four more broadly defined genres. In this case, we examined only the four broad genres, namely:

- Fiction
- Learned (i.e. academic prose)
- Press (news, editorials, reviews)
- General prose (non-academic non-fiction)

We defined subcorpora of each of these divisions of the NNC-CS, and ran a key tags analysis, the results of which are shown in Table 1. [4] For each broad genre, the sixteen most positively key tags (significantly *more* frequent in that subcorpus) and the sixteen most negatively key tags (significantly *less* frequent in that subcorpus).

Key tags in "fiction"		Key tags in "general prose"		Key tags in "learned"		Key tags in "press"	
More common	Less common	More common	Less common	More common	Less common	More common	Less common
PMX	JX	MM	VVYN1F	JX	PMX	NP	PMX
TT	NN	FS	TT	IKO	VVMX1	JX	VVMX1
VVMX1	NP	NN	IKF	CC	PMXKM	FB	DDX
VVYN1F	IKO	IH	VDX	FS	PTM	VE	FS
IKF	IH	JX	VVYM1F	NN	VQ	IKO	TT
VVYM1F	FS	CC	IE	IKM	VVTX2	IKM	VQ
PTM	CC	FZ	RK	FZ	VDX	NN	IKF
VDX	II	II	VE	II	IE	PXH	RR
VQ	IKM	VN	DDX	MOX	JM	IH	VVYN1F
DDX	MM	VVYX2	NP	CBS	PXH	CSA	PMXKM
RK	VN	FO	PMX	FU	MLX	IE	VVYX2
VVYN1	FB	FF	PTM	MOM	VCM	VI	VVYN1
PMXKM	FZ	PMXKO	VVYN1	IKX	PTH	VN	PTH
UU	FO	VDO	IA	JT	UU	II	VVTX2
PTH	MOX	DJX	PTN		RK		JM
VVTX2	MOM	DGM	VVMX1		TT		VCM

Table 1. Key POS tags analysis across genres of the NNC-CS.

While a full analysis of this data lies beyond the scope of this working paper, certain trends are immediately evident. The positively key tags for fiction include several different verbal tags (V-), pronominal tags (P- and D-), and tags associated with animate referents (the tags ending in –F for feminine: Nepali gender agreement only operates with animate entities). By contrast, adjectives (JX), nouns, (N-), numerals (M-) and adpositions (I-) – all categories associated with complex lexical noun phrases – are negatively key. So fiction in Nepali is characterised, quantitatively, by a high frequency of verbs and pronominal noun phrase and a low frequency of lexical

[4] The software used to carry out all the searches described in this WP was *CQPweb*. See Hardie, forthcoming. Subcorpus definition and key tag searches are built-in functions of CQPweb. See also http://www.ling.lancs.ac.uk/activities/713/ .

noun phrases. By contrast, learned prose s characterised by more or less the exact opposite. Most of the negatively key tags are verbs and pronouns, and most of the positively tags represent adjectives, nouns, adpositions, and numerals.

The press category also has many key tags linked to lexical noun phrases – especially nouns and adpositions – although, in contrast to learned prose, not numerals. Although many of the negatively key tags represent verbs, there are also positively key verb tags – mostly representing infinitive and participle forms (VE, VI, VN). So rather than a significant lack of verbs, there seems to be a preference for particular *types* of verb forms. Again, many pronoun categories are negatively key. Finally, general prose presents a mixed picture. There are noun tags, adposition tags, verb tags and pronoun tags in both the positively and negatively key lists for this genre.

To summarise this picture, we can present the relationship among the genres as a continuum, as follows:

Verbs and pronouns prominent	Fiction > General prose > Press > Learned	*Nouns, adjectives and adpositions prominent*

It is interesting that these results are generally in line with those of Rayson et al. (2002) on English texts. Working with two rather different categories of written text, derived from the structure of the British National Corpus, this study also observed that pronouns and verbs are more frequent in *imaginative writing* (i.e. fiction) than *informative writing* (i.e. non-fiction), whereas nouns, adjectives, and adpositions are more frequent in informative writing than imaginative writing (Rayson et al. 2002: 301-304). This suggests a cross-linguistically valid pattern of variation.

The hypothesis that will be argued here is that this pattern is linked to *narrative*, as one way to characterise this continuum of text types is as a continuum between, on the one hand, narrative content, and on the other, abstract informational content. Fiction has a very strong narrative concern. By contrast, academic prose is focussed on the conveyance of formalised, often abstract information, rather than narrative. This hypothesis is based on a comparison of these findings to those of Biber (1988; see esp. pp. 104-120). Using a factor analysis of multiple linguistic features across a variety of text-types, Biber characterises narrativity as one of the main dimensions of textual variation, along with others, noting that, as in Biber (1986), his analysis identifies

> three major dimensions that mark (1) interactive, involved discourse versus edited, informational discourse; (2) formal, abstract information versus non-abstract types of information; and (3) reported, narrative discourse versus non-narrative types of discourse
> (Biber 1988:119)

Linguistically, the key tags that have been discussed here represent features that Biber locates on different dimensions. For instance, some types of verbs are characteristic of involved (rather than informational) discourse, whereas others are associated with narrative (rather than non-narrative) discourse. By contrast, nouns, adpositions and attributive adjectives are characteristic of informational discourse. However, using the data presented above, and without further work to tease apart the sub-genres within

the NNC-CS, it is not possible to differentiate multiple dimensions of variation using just the key tag data. Rather, at this level of granularity only a single "dimension" can be identified – that of narrative versus abstract or informational discourse. This is the trend identified above.

It is at this point that the Bandhu Collection becomes invaluable to the analysis.

- Consisting largely of spoken narratives, the Bandhu Collection offers the means to test the hypothesis of narrativity as a factor.
- By looking at the incidence of *other* linguistic features within the Bandhu Collection (as opposed to the NNC-CS), their association, or lack thereof, with the narrativity parameter may be ascertained.

In the present working paper, the first of these points will be addressed briefly, once more by means of key POS tag analysis. Finally, the second point will be addressed with reference to the link to narrativity of a single feature, namely progressive aspect, as manifest in two auxiliary verbal markers in Nepali.

Part-of-speech frequencies in the Bandhu Collection

A list of key POS tags for the Bandhu Collection as opposed to the *entire* NNC-CS was generated; is given as the Appendix. Note that this means that the Collection is being contrast to *all* the categories of the NNC-CS discussed above. If the hypotheses about narrative vs. abstract informational discourse discussed above hold true, then we would expect the trends evident in the fiction subcorpus to be even more evident in the Bandhu Collection, even when compared to a point of reference that includes the fiction genre.

This prediction is generally fulfilled by the data. Of the 29 positively key tags identified for the Bandhu Collection, 14 are verb tags, and a further seven are pronoun tags. By contrast the majority of the negatively key tags represent adjectives, nouns, adpositions and numerals. This would seem to confirm the hypothesis that the Bandhu Collection is more strongly marked by the trends identified in the NNC-CS fiction texts. However, there is some counter-evidence, namely that some verb tags and one pronoun tag are negatively key. This may require further investigation to be elucidated.[5] However, on the whole the key tags of the Bandhu Collection provide strong support for the notion that the Collection can be used as a means to triangulate the impact of narrativity on different linguistic features within the NNC-CS and between the NNC-CSand the Collection itself.

Progressive aspect markers in Nepali

The grammar of the Nepali verb is characterised by a very high degree of compounding of finite auxiliary verbs with non-finite verbs to create complex tense-

[5] The positive keyness of the JF tag in the Bandhu Collection is not counter to the narrativity hypothesis, because it is a feminine tag: see the discussion of the relationship between frequency of feminine inflections and animate reference (and thus with fiction and narrativity) above.

aspect-mood forms. Progressive aspect is indicated, in particular, by two markers: *dai* and *rah*.

Other forms, such as the root of the verb *rākhnu* "put", may also be used to indicate progressive meaning. However, these are not as central to the verbal paradigm as *rah* and *dai* and will not be considered further here.

Dai is, in origin, a participle inflection. It is described by Acharya (1991: 147) as a conjunctive participle and translates *gardai* as "doing". *rah* is the root of the still-extant independent verb *rahnu*, meaning "stay/remain". Its progressive usage appears to be a development of its lexical meaning via grammaticalisation.

Both can indicate progressive aspect, with only a very slight if any difference between them. It is possible for both to occur on the same verb. Furthermore, *dai* also occurs in many verbs not as a marker of aspect, but rather as part of the negative inflection (compare *gar-cha*, "he does", and *gar-dai-na*, "he does not").

Both forms present practical difficulties in searching for. In the first place, to identify instances of *rah* as an aspect marker, it is necessary to first search for all instances of verb-forms containing *rah*, and then to subtract those where *rah* is the lexical verb.[6] *Rah* is considered an auxiliary when it is compounded after another verb form. However, (grammatically) compound words have variant orthography in Nepali, so such a compound might be represented in the NNC as one or two orthographic tokens. So the number of instances where (a) *rah* is word-initial *and* (b) the preceding word is not tagged as a verb was calculated, and this subtracted from the overall frequency of verbs containing *rah*.

A similar procedure was be undertaken for *dai*. First a count of all verbs containing this syllable was obtained. Then, the number of instances where this is not an aspect marker was calculated by searching for verbs containing *daina*.

For technical reasons of text encoding, the searches in the Bandhu Collection could not be accomplished by the same means, and will be discussed at the appropriate points below.

Progressive markers *rah* and *dai* in the Bandhu Collection and the NNC-CS

The frequencies calculated for the *rah* in the NNC-CS were, once again, calculated according to the broad genres used in the key tag analysis above. Table 2 shows the frequencies of *rah* overall, as an aspect marker, and as a main verb, calculated as described above.

[6] It is of course questionable whether a verb with the lexical meaning of *stay* could not be considered an indicator of aspect in the widest sense. A similar consideration might apply to the use of *dai* in negative forms. However, a full discussion of this issue lies beyond the scope of this paper.

Genre	**Overall frequency of *rah***		**Frequency of *rah* as aspect marker**		**Frequency of *rah* as main verb**	
	Raw freq	*Per 1000*	*Raw freq*	*Per 1000*	*Raw freq*	*Per 1000*
Fiction	1941	6.169	1548	4.920	393	1.249
General prose	2562	5.358	1291	2.700	1271	2.658
Learned	789	4.969	233	1.467	556	3.501
Press	957	5.575	472	2.750	485	2.825

Table 2. Absolute and relative frequencies of *rah* in the NNC-CS

The interesting feature that emerges here is that the genres differ, not only in the frequency of *rah* as an aspect marker, but also in their use of *rah* as a main verb. In general prose and press, there is a roughly even split: in fiction there are far more auxiliary than main forms and in learned texts there are far more main than auxiliary forms. The overall frequency of *rah* is also highest in fiction and lowest in learned prose. Again we see a continuum of this sort:

Rah *as aspect marker is prominent*	Fiction > (General prose & Press) > Learned	Rah *as main verb is prominent*

Therefore, it may be hypothesised that *rah*-as-aspect-marker is a feature associated with the dimension of variation established above. If so, we would expect it – in keeping with the key tag results – to be even more commonly an aspect marker in the spoken narratives of the Bandhu Collection. Unfortunately, the technical issues alluded to above currently impede a search of the Bandhu Collection of the type shown in Table 2. Rather, it was only possible to search for the overall frequency of *rah*. However, as noted above, there is a correlation across the NNC-CS genres between overall frequency and prominence of aspect marker over main verb, so this functions as an appropriate proxy variable. The frequency of *rah* overall in the Bandhu Collection is approximately 377, i.e. about 7.6 per thousand words. This is indeed higher than the overall frequency in fiction (6.2 per thousand words). This is not fully convincing as evidence for the association of rah-as-aspect-marker with narrativity in Nepali, but it is suggestive. Further work is clearly needed here to repeat the searches of the Collection in a more rigorous way and to confirm this preliminary conclusion.

The picture that emerges with regard to *dai* (see Table 3) has both similarities to, and differences from, *rah*. The *overall* frequency of *dai* follows the same pattern as the overall frequency of *rah*. Furthermore, the pattern of *dai*-as-aspect-marker follows the pattern of *rah* as aspect marker. That is, in both cases the feature in question is most frequent in fiction, least frequent in learned prose, with general prose and press text having middle values (here, press is slightly above prose). The first difference here is that, in the case *dai*, both overall counts and count-as-aspect-marker are *much* greater than in the other three genres. The other difference is that there is *not* a complementary pattern of *dai* in its other use. Instead, all the genres but the press texts are roughly similar, and the press texts are an outlier. We might tentatively conclude that *dai*-as-aspect-marker and *dai*-as-part-of-negative are more differentiated as separate features than are the main-verb and aspect-marker uses of *rah*.

Genre	Overall frequency of *dai*		Frequency of *dai* as aspect marker		Frequency of *dai* as part of negative	
	Raw freq	*Per 1000*	*Raw freq*	*Per 1000*	*Raw freq*	*Per 1000*
Fiction	3209	10.200	2502	7.953	707	2.247
General prose	2679	5.602	1586	3.317	1093	2.285
Learned	783	4.931	421	2.651	362	2.279
Press	1078	6.280	829	4.830	249	1.450

Table 3. Absolute and relative frequencies of *dai* in the NNC-CS

However, be this as it may, the general cline observed from fiction to general prose/press to learned prose is preserved for *dai*-as-aspect-marker, and we would thus predict that it would be as frequent or more frequent than in fiction in the Bandhu Collection. While again encoding issues impede the production of a directly comparable figure for the Collection, *dai* appears to occur around 350 times overall (7.1 per thousand), of which 250 are aspect markers (5.1 per thousand). This runs contrary to the prediction: the values for the Bandhu Collection are *between* those for fiction and press texts. While this is an interesting variation, it is still consistent with an interpretation of progressive aspect – both *rah* and *dai* – as a linguistic feature associated with narrativity.

Conclusion

This working paper has laid out some initial analysis in support of the following claims.

- There exists a quantitative cline across the grammatical features of the broad genres of the NNC-CS, from learned prose, to general prose and press discourse, to fiction
- High frequency of nouns, adpositions, adjectives and numerals characterises the former end of this cline; high frequency of verbs and pronouns characterise the latter end
- Narrativity is a significant factor in this cline: texts at the "learned" end are non-narrative, abstract and informational, whereas text types at the "fiction" end are narrative and involved (to use Biber's 1988 terms).
- In Nepali, progressive aspect is associated with the function of narrativity. Progressive aspect markers increase in frequency in more narrative-like types of text.
- The Bandhu Collection, as a body of spoken, primarily narrative texts, is located near or beyond the "fiction" end of this scale, depending on which linguistic feature is being analysed and thus allows us to reconfirm and triangulate the effect of narrativity in the genres of the NNC.

This research is still in its initial stages and much further work remains to be done. In particular, technical advances in the encoding of the Bandhu Collection will allow us to run more rigorously compatible searches on the data alongside the NNC-CS. They will also allow us to separate out the minority of non-narrative (dialogue) data from the Bandhu Collection to create a more monotypic block of spoken narrative data.

However, the consistency of the results obtained with even a limited methodology is encouraging, and highly suggestive.

To expand the work, we will look at other verbal markers than *rah* and *dai* – both aspect markers and other verbal categories. For example, finite forms of the verb *hunu*, "be" (such as *cha*, "is") are used in Nepali as compounded auxiliary verbs to form finite tenses of other main verbs. The similarity or difference of the distribution of auxiliary *hunu* – in its various forms – will shed further light on the linguistic features associated with text type variation in Nepali.

Thinking more broadly, a thorough analysis of aspect marker frequency across genres would enable the Nepali data to be compared to the analysis of aspect in Mandarin of Xiao and McEnery (2004). Similarities and differences between aspect in Mandarin, a highly isolating language, and Nepali, which as noted above has a highly complicated system of agglutinative/compounding verb inflection, cannot but be of interest to researchers on both languages. However, in terms of genre/text type in particular, a cross-linguistic comparison with English may also be illuminating here. In Biber's (1988) analysis, different verbal features are associated with different dimensions of variation. We may anticipate that, when this research is sufficiently advanced to differentiate more than a single dimension of variation, a similar phenomenon may emerge for Nepali. Comparing which verbal features are associate with particular dimensions in the two languages may shed light on the correspondences between two drastically different, but both highly complex, tense-aspect-modality systems.

References

Acharya, J (1991) *A descriptive grammar of Nepali*. Washington, D.C.: Georgetown University Press.

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D (1986) Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62: 384-414.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, Volume 19, number 1, pp. 61-74.

Francis, W.N. and H. Kučera (1979). Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers. Revised and amplified version. Available at http://khnt.hit.uib.no/icame/manuals/brown/index.htm .

Hardie, A (2007) Part-of-speech ratios in English corpora. *International Journal of Corpus Linguistics* 12(1): 55-81.

Hardie, A (*forthcoming*). "CQPweb".

Hardie, A, Lohani, R, Regmi, B and Yadava, Y (2005) Categorisation for automated morphosyntactic analysis of Nepali: introducing the Nelralec Tagset (NT-01). Nelralec/Bhasha Sanchar Working Paper 2. Available at http://bhashasanchar.org/pdfs/nelralec-wp-tagset.pdf .

Hardie, A, Lohani, RR, Regmi, BN and Yadava, YP (*forthcoming*). A morphosyntactic categorisation scheme for the automated analysis of Nepali.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language* 70(2): 331-339.

Hundt, M., A. Sand, and R. Siemund (1998). Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB'). Englisches Seminar, Albert-Ludwigs-Universität Freiburg. Available at http://khnt.hit.uib.no/icame/manuals/flob/index.htm .

Johansson, S., G. Leech, and H. Goodluck (1978). Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. Oslo: Department of English, University of Oslo. Available at http://khnt.hit.uib.no/icame/manuals/lob/index.htm .

Mair, C., M. Hundt, G. Leech, and N. Smith (2002). Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. In: International Journal of Corpus Linguistics 7(2): 245-264.

Rayson, P., A. Wilson, and G. Leech (2002). Grammatical word class variation within the British National Corpus sampler. In: P. Peters, P. Collins, and A. Smith (eds.) New frontiers of corpus research: Papers from the Twenty First International Conference on English Language Research on Computerized Corpora, Sydney 2000. Amsterdam: Rodopi.

Xiao, Z. & A. McEnery (2004) *Aspect in Mandarin Chinese: A corpus-based study* . Amsterdam : John Benjamins.

Appendix

Key POS tag table for the Bandhu Collection vs. the NNC-CS. + / – indicates a positively/negatively key tag. For a key to tags, see Hardie et al. (2005, forthcoming).

Number	Part-of-speech tag	Freq (BC)	Freq (CS)	+/-	LL
1	JX	643	67427	–	2603.28
2	RD	1089	5745	+	1397.11
3	VQ	2131	19050	+	1309.13
4	NP	206	24926	–	1061.57
5	UU	300	496	+	888
6	TT	2464	28393	+	868.78
7	YF	4200	57932	+	848.37
8	IKO	213	20963	–	771.03
9	VVYN1	3777	52465	+	743.3
10	IKM	1634	68067	–	715.41
11	YQ	1247	13756	+	491.1
12	PTH	244	936	+	423.15
13	CC	729	29812	–	297.93
14	IH	385	18974	–	293.36
15	VVYX2	1151	14857	+	292.01
16	II	2381	75122	–	275.61
17	VDO	445	4057	+	263.59
18	MM	578	23399	–	227.28
19	VCH	83	342	+	135.38
20	VE	1063	33787	–	129
21	JF	86	394	+	127.28
22	VCN	57	181	+	114.55
23	YM	1729	30060	+	108.54
24	DDM	153	1261	+	108.32
25	VVTN1	53	165	+	108.05
26	VN	1113	18494	+	94.03
27	VVMX1	360	4804	+	81.79
28	VCM	108	858	+	81.18
29	RR	530	17511	–	81.17
30	NN	12174	300322	–	80.52
31	MLO	28	2391	–	77.86
32	PMX	491	7276	+	74.47
33	DDX	1662	30351	+	70.73
34	VDM	67	476	+	59.57
35	FB	80	3895	–	58.78
36	JM	178	6640	–	50.01
37	PTM	120	1324	+	47.24
38	PTNKF	8	5	+	33.77
39	VVTX2	69	714	+	31.55
40	PTNKO	7	7	+	25.53
41	PRFKM	66	2655	–	25.33
42	VS	22	1223	–	23.46
43	VVTN1F	10	32	+	19.98
44	VDX	294	5096	+	18.86
45	FZ	50	565	+	18.54
46	VVMX2	26	1241	–	17.98