Collocational Properties of Adpositions in Nepali and English

Andrew Hardie[1]

1. Introduction

This paper presents the results of a preliminary investigation into the use of statistical collocation as a method for discovering characteristic properties of grammatical classes (i.e. part-of-speech categories such as noun, adjective, verb, and so on). In particular, the focus of the current investigation is the collocational patterns found around adpositions in English and Nepali.

The claim that will be made on the basis of the findings presented here is twofold. Firstly, two primary collocational patterns characterise the category of adposition in English and Nepali. Secondly, these patterns are observable on different adpositions to different degrees. This claim is illustrated with reference to detailed discussion of a subset of the most frequent adpositions in English and Nepali.[2]

The paper is organised as follows. §2 presents an overview of the theoretical background to the "quantitative-distributional" method followed here. The method itself, and the data to which it will here be applied, are then detailed (§3). The collocations found for the adpositions under investigation, and a discussion of the patterns that can be detected among those collocations, are presented in §4 (English data) and §5 (Nepali data). The paper concludes with a discussion of these findings (§6) and some comments on possible further developments of this research (§7).

2. Background

A number of different theories of language agree, in broad terms, that grammatical categories – such as noun, verb, and adposition – are not basic components of the language system, but are, rather, phenomena that emerge from the distribution of lexical items. For instance, Hoey's theory of Lexical Priming argues that words are psychologically "primed" to co-occur with other words, and that part-of-speech categories to which words are assigned are actually labels for the combinations of primings that are most characteristic of those words (see Hoey 2005: 7-8, 154-155). So category membership is determined by co-occurrence, i.e. by distributional properties of the words in the category. Crucially, these co-occurrence relations can be discovered by the examination of collocations, colligations and semantic associations in text corpora.

From this perspective, it might be asked whether it is possible to characterize a part-of-speech category entirely in terms of the collocational patterns that can be observed around the words that occur within that category. This paper represents an attempt to do so for the category of adpositions, using statistical collocation as a tool for accessing the distributional properties of particular adpositions. Since adpositions

[1] Department of Linguistics and English Language, Lancaster University
e-mail: a.hardie@lancaster.ac.uk

[2] The Nepali data reported here is analysed in greater depth in Hardie (forthcoming).

are (a) relatively few in number and (b) individually frequent in running text, they are a suitable initial category to attempt to characterise in these terms.

As this approach involves looking at distribution in a fundamentally quantitative rather than qualitative way, it can be referred to as a "quantitative-distributional" method. In broad terms, statistical significance is here treated as a proxy variable for conceptual salience. If a collocate, or a pattern instantiated across several collocates, scores highly according to the collocation statistic (in this case Z-score: see below), then it is remarkably frequent in the vicinity of the node. Assuming – as a range of models of language acquisition such as Tomasello (2003) do assume – that semantic and morphosyntactic properties of words are learnt from exposure to multiple examples of their usage in context, the remarkably frequent aspects of a word's context (i.e. the high-scoring collocations) will therefore be conceptually prominent aspects of a speaker's knowledge of that word. So characterising a grammatical category using patterns among their high-scoring collocates is an investigation by proxy of the conceptually most salient features of that category's distributional behaviour.

To ensure cross-linguistic validity of the findings, this analysis is applied here to two languages: English and Nepali. This pair of languages is an interesting testing ground for this analysis for a number of reasons. While English has prepositions, with a small number of arguable postpositions such as *'s* and *ago*, Nepali has only postpositions. This is in keeping with their basic word order: English is SVO and Nepali is SOV, language types which typically have prepositions and postpositions respectively (Greenberg 1963). However, as both are Indo-European languages, they are otherwise similar on many points. So English and Nepali is an appropriate language pair to investigate the cross-linguistic dimension of collocation-based characterisation of adpositions as a category.

Some previous work has used corpus data to analyse particular English prepositions. For example, Sinclair (1991) uses corpora to investigate the wide range of functions of *of*, and Kennedy (1991) uses corpus data to delineate the difference between the semantically similar prepositions *between* and *through*. However, none of this work has used collocations derived using significance statistics, which is the basis of the "quantitative-distributional" approach applied here. Given that no corpus data was previously available for Nepali, there has not previously been any corpus-based analysis of Nepali postpositions.

3. Data and method

Four corpora were used to derive the collocation statistics discussed in this paper. For English, the FLOB corpus (Hundt *et al*., 1998) was the primary data; but the spoken and written components of the BNC Sampler[3] were subjected to the same analysis, to provide a point of comparison. Each of these corpora consists of one million words. Given the very great frequency of the adpositions discussed here, one million words of running text is more than enough to provide statistically significant collocations. For Nepali, a subset of the Nepali National Corpus (NNC) was used. The NNC is a large corpus currently under development.[4] One component, the "Core Sample", is a Nepali match for FLOB (i.e. texts sampled from the early 1990s). As such, it follows

[3] See http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products#sampler
[4] See http://www.bhashasanchar.org/new/

the Brown Corpus sampling frame, on which FLOB is also based. The data utilised here consists of the first part of that core sample to be assembled in electronic form. More precisely, it consists of 165 text samples of 2000 words each, out of the 500 texts in the target sampling frame.[5] While the relative proportions of each genre within the current dataset are not the same as they will be in the finished NNC Core Sample (and as they are in the matching FLOB corpus), the genre make-up of FLOB and the Nepali dataset is nonetheless similar, more similar for instance than FLOB and the written BNC sampler.[6]

Although only about one third of the size of English datasets, the subsection of the NNC used here is still large enough to yield statistically valid collocations for items as frequent as Nepali postpositions. As the NNC was incomplete at the time this research was undertaken, one obvious extension of this investigation will be to repeat the analysis on larger samples of Nepali text such as the completed NNC core sample. Based on the similarities found between the different English corpora (FLOB, written BNC sampler and spoken BNC sampler: see discussion below), it is anticipated that changing the dataset may affect the precise lexical items that occur as collocates but will not affect the patterns that exist across the different collocates. Of course, this remains to be established empirically.

For each of the words under investigation, collocation lists were created using the Xaira software.[7] Xaira can score collocates by Z-score or by Mutual Information; for this investigation the Z-score was used, as MI is known to overestimate the importance of low-frequency collocates (see, for example, the demonstration of this by Baker, 2006: 102). The search was run for collocates within a window of two words to the left and the right of the adposition. It might be argued that this is a sub-optimal approach, given that it is known that adpositions are grammatically linked to their NP complements: thus the most meaningful results should in theory be gained from the analysis of the material that appears *after* prepositions and *before* postpositions. This argument fails on two grounds. Firstly, it relies on prior theoretical knowledge of the nature of adpositions, whereas the aim of this study is to investigate whether adpositions can be characterised based on statistical collocation *alone* – which requires the exclusion, as far as is possible, of prior knowledge about what directions of co-occurrence are or are not relevant. Secondly, empirical evidence indicates that there *do* exist significant and relevant collocates which occur before prepositions or after postpositions, which *must* be taken into account in an adequate characterisation of the collocational properties of adpositions. For instance, Sinclair's (1991: 81-98) study of English *of* establishes a typology of functions based on both preceding and following lexical items. In the results below, there is at least one instance of a lexical item occurring in significant patterns both before and after the same postposition (*ādhāra mā* and *mā ādhārita*: see §5).

The collocation lists were restricted to the twenty items with the highest Z-scores, partly in the interests of making the data manageable, but also to focus attention tightly on the most significant relationships in the corpus. These collocation lists were then examined for any patterns of similarity or contrast among the different postpositions. It was not considered necessary that the analysis should account for every single collocate. Given the relatively small size of the different datasets, *individual* collocates can easily be caused by the prominence of a particular word in a

[5] For a more detailed description of the genres in the sampling frame, see Hundt *et al*. (1993).
[6] See http://www.comp.lancs.ac.uk/ucrel/bnc2sampler/sampler.htm
[7] See http://www.oucs.ox.ac.uk/rts/xaira/

single text sample. Rather, the aim was to identify general patterns characterising the *majority* of the collocations of the word under investigation.

A technical problem in analysing the Nepali postpositions is that, in the Nepali writing system, postpositions are often (but not always) attached to the word that they follow – much as English *'s* is attached to the word it follows. It was therefore necessary to split off the Nepali postpositions by using a tokeniser[8] to preprocess the text (in the same way that English *'s* is split off from the word it follows by most English tokenisers). However, the splitting process produces some false positives. For instance, one of the Nepali postpositions discussed here is *mā*. The word *paramātmā* ("God") ends in *mā*, but it is not a postposition – it is simply part of the base form of the noun. If *mā* is split off here and counted as a postposition, this is inaccurate and could skew the results – especially since *paramāt* will only ever occur directly before *mā*, thus falsely indicating a strong collocational link. A list of such exceptions was built in to the tokeniser, which could then avoid making the mistake. However, this list can never be fully complete, and as a result the collocation lists derived from the data did sometimes include non-word elements such as *paramāt*. Where this occurred, the list was manually edited to remove these errors prior to analysis.

A total of nine prepositions are sufficiently frequent in English to yield useful statistical collocation data in these datasets: *of, to, in for, on, with, by, at* and *from*. Since an extensive consideration of all these words requires more space than is available here, three prepositions representative of the different patterns observed (*in, with* and *by*) are analysed in detail below. Since the Nepali dataset was smaller, only five postpositions were sufficiently frequent for this analysis: *ko / kā / kī*, *mā*, *le*, *lāī*, and *bāṭa*.[9] The latter four are analysed in this paper.

All the items whose collocations are examined in this paper are extremely frequent. It is, therefore, necessary to eliminate the possibility that the patterns observed are not due solely to high frequency. This is relatively simple to do by looking at the collocation lists for some other extremely frequent words. The words selected for this purpose were *and* and *he* in English, and the translation-equivalent words (*ra* and *yasa*) in Nepali. Space restrictions prohibit the presentation of this data here; it will suffice to note that the patterns described in the following section are *not* observed around these other high-frequency types. We can therefore be confident that the patterns discussed below are not an epiphenomemon of extreme high frequency and are thus meaningful for the characterisation of adpositions.

4. Patterns in the English data

As noted above, the primary claim of this paper is that there are particular collocational patterns that characterise adpositions as a category. In this section, evidence for this proposal will be presented from the collocates of three prepositions in English; the following section will present evidence for a similar pattern in Nepali.

The two main patterns that will be proposed as diagnostic of prepositions in English are both found around the word *in*; hence, this will be the first adposition to be discussed. The collocations of *in* in the FLOB corpus are shown in Table 1.

[8] The software used for this was *Unitoken*, a component of the *Unitag* language-independent part-of-speech tagging architecture (see Hardie 2004, 2005).

[9] See §5 for translations. Nepali is written in the Devanagari script; however all Nepali words in this paper are given in transliterated form.

	FLOB		
No.	Collocate	Frequency	Score
1	The	7008	40.5
2	Fact [B]	193	31.5
3	case [B]	194	28.6
4	Britain [A]	119	21
5	Cases [B]	87	20.6
6	Ways	82	20.3
7	Early	123	20.3
8	Interested [C]	57	20
9	England [A]	110	20
10	Detail [B]	53	19.5
11	Involved [C]	83	19.5
12	London [A]	137	18.8
13	Interest [C]	95	18.7
14	Context [B]	55	18.5
15	europe [A]	82	18.4
16	scotland [A]	53	18
17	france [A]	60	17.4
18	america [A]	64	17.3
19	Middle	58	16.7
20	Differences [C]	55	16.4

Table 1: Collocations of *in* in FLOB.

A number of patterns can be detected here.[10] One is the presence of a large number of place nouns (countries, cities). There are also nouns that form phrases where *in* has a metaphorical meaning (*context, fact, case*). These patterns are labelled [A] and [B] respectively in Table 1. The other noticeable pattern, embodied in the collocates labelled [C], is that there are collocational links to lexical items for which the preposition *in* functions as a subcategoriser – that is, in these cases *in* is a linking element between its collocate and another nominal which refers to a participant in some state-of-affairs referred to by the collocate. The collocates in question are primarily verbs (*interested in X, involved in X*), but sometimes nouns (*interest in X, differences in X*).

It would be difficult to argue that these patterns are meaningful solely on the basis of the data presented so far. However, their existence is confirmed (i) by the appearance of the same patterns, to a greater or lesser extent, on the other very frequent prepositions (discussed below); and (ii) by the appearance of the same patterns, but *not* all the same lexical items, in the BNC Sampler datasets – see Table 2. Note in particular that a wider range of collocates for which *in* functions as a subcategoriser can be seen here. Also note that in the spoken data, there are fewer "city" or "country" nouns of place, and more nouns of what we might describe as

[10] In many cases it was necessary to examine concordances of the node word and collocate co-occurring to identify patterns with certainty. In the interest of brevity, this concordance data is not presented here.

"personal" place: *bed*, *flats*, *house*, *room*. In other words, the same semantic pattern (nouns of place) can be detected across datasets, even though they are different *types* of nouns of place, presumably as a result of text type differences. There are also, in these datasets, hints of a pattern of words relating to time (*1987, early, recent, morning*) although there are insufficient examples of this for the pattern to be wholly clear.

	BNC Sampler (written)			**BNC Sampler (spoken)**		
No.	Collocate	Frequency	Score	Collocate	Frequency	Score
1	the	7165	38.7	the	3976	58.1
2	case [B]	171	28	fact [B]	234	54.8
3	europe [A]	132	22.8	n	241	41.6
4	1987	64	21.4	put [C]	339	29.5
5	britain [A]	125	21.3	middle	71	29
6	increase [C]	83	20.4	involved [C]	59	25.1
7	washington [A]	63	20.4	interested [C]	55	24.9
8	london [A]	131	20.2	case [B]	81	23.1
9	shown [C]	102	20.1	scotland	45	21.3
10	fact [B]	109	19.3	favour	26	21.2
11	involved [C]	74	18.7	nineteen	88	20.6
12	cases [B]	62	18.5	morning	113	19.8
13	uk [A]	88	18.5	country [A]	60	19.1
14	interested [C]	41	18.5	london [A]	50	18.9
15	early	118	17.4	minute	66	18.7
16	recent	77	17.2	bed [A]	73	18.5
17	practice [B]	54	16.3	living [C]	52	18.5
18	place [A]	126	16.3	flats [A]	54	18.4
19	vienna [A]	24	15.8	house [A]	97	17.9
20	paris [A]	41	15.4	room [A]	77	17.4

Table 2: Collocations of *in* in the BNC Sampler.

The pattern labelled above as [C] – where *in* functions as a subcategoriser for the collocate – is found with several other prepositions. In some cases, it is much more dominant than it is for *in*. An example of this is seen upon examining the collocations of *with* (see Table 3).

There are many more verbs on this list, and also numerous nouns and adjectives for which *with* is a subcategoriser: *links, familiar, contact, consistent, relationship, conformance, wrong*. In fact, the only top-twenty collocates which are not covered by this pattern are *a* and *together*. This pattern is borne out by the collocation lists from the BNC Sampler, although in the case of the spoken Sampler there are slightly more collocates that do not fit the pattern (*happy, the, up, wi, a*); see Table A1 in the Appendix.

	FLOB		
No.	Collocate	Frequency	Score
1	associated [C]	67	44.8
2	compared [C]	52	34.9
3	deal [C]	59	31.7
4	filled [C]	42	29.9
5	cope [C]	29	29.3
6	dealing [C]	27	28.9
7	concerned [C]	58	27.2
8	dealt [C]	22	27
9	contact [C]	43	25.2
10	consistent [C]	26	23.4
11	a	1077	22.8
12	connected [C]	19	18.9
13	compete [C]	16	18.3
14	coupled [C]	12	18.1
15	together	61	18
16	links [C]	20	17.5
17	comply [C]	8	16.7
18	relationship [C]	37	16.7
19	deals [C]	15	16.5
20	equipped [C]	12	15.8

Table 3: Collocations of with in FLOB.

By contrast, let us now consider *at* (see Table 4). *At* functions as a subcategoriser for a number of verbal collocates, indicated with [C] in Table 4, in keeping with the labels used above. However, in this case, the verbs it subcategorises for are semantically limited: all refer to various forms of "looking" (*stared, looked, glanced*). One verbal collocate (*educated*) in the written BNC Sampler (see Table A2 in the Appendix) is an exception to this, but otherwise the pattern holds. So this pattern is arguably somewhat less prominent and general than it was for *in*, and much less prominent than it was for *with*.

The other pattern observed for *in* – collocation with nouns that the preposition does not function as a subcategoriser for – was not observed for *with* but is evident in the data for *at*. The nouns observed as collocates of *in* fell into two groups, nouns of place [A], and nouns forming phrases where the postposition is metaphorical [B]. In the data on *at*, a pattern related to the former pattern emerges clearly: however in this case the nouns are more often stereotypical nouns of time than nouns of place. The nouns of place that are significant are *home* and *school*; intuition suggests, and examination of the concordances confirms, that we are here seeing the impact of the idiomatic phrases *at home* and *at school*. The fact that (at least) two such idioms exist suggests that this is a relevant collocational pattern of *at*, rather than something unique and exceptional that should be excluded from the analysis. We can also count *same* as part of this pattern, as the concordance suggests that its appearance as a high collocate is due to the phrase *at the same time*. The nouns of time and place are

marked as [A] in Table 4, in parallel to the labelling adopted above. The same two patterns can be seen in the BNC Sampler data for *at* (see Table A2 in the Appendix).

	FLOB		
No.	Collocate	Frequency	Score
1	Stared [C]	45	41.6
2	look [C]	118	41.5
3	looked [C]	106	39.2
4	aimed [C]	30	36.7
5	glanced [C]	32	35.2
6	Time [A]	184	33
7	end [A]	96	31.3
8	home [A]	91	27.2
9	the	1900	26.7
10	looking [C]	54	25.4
11	staring [C]	15	25
12	same [A]	88	24.5
13	intervals [A]	13	24.1
14	Expense	17	22.8
15	school [A]	56	22.6
16	moment [A]	46	21.9
17	beginning [A]	32	20.6
18	gazing [C]	10	19.7
19	Temperatures	14	19.6
20	outset [A]	9	19

Table 4: Collocations of at in FLOB.

In summary, then, this examination of three of the nine most common prepositions in English (*at, in* and *with*) suggests that there are two or possibly three main patterns:

- The pattern labelled [C] above, where the collocate is a word – most often a verb but also possibly a noun or adjective – for which the preposition functions as a subcategoriser.
- The pattern labelled [A] above, where the collocate is a noun whose semantics are coherent with those of the preposition (exemplified here by nouns of place and time with *in* and *at*)
- The pattern labelled [B] above, where the collocate is a noun which, with the preposition, forms a phrase in which the preposition has metaphorical meaning (e.g. *fact, case* as collocates of *in*).

It is not entirely clear that [B] is in fact a separate pattern from [A]. It could be simply another form of semantic coherence between the preposition – here in its metaphorical sense – and the collocate. Further investigation of this point is probably

warranted. For now, in a spirit of parsimony, it will be assumed for current purposes that [B] can be subsumed within [A] and, therefore, that it is the two patterns labelled [A] and [C] that are characteristic of the collocational behaviour of prepositions.

However, as the comparison of *in, with* and *at* above has demonstrated, these two patterns are not *equally* characteristic of all prepositions. English prepositions vary in terms of how prominent the patterns are in the most significant collocations of that preposition. *In* shows both patterns, whereas *with* shows only the "subcategoriser" pattern [C], and *at* shows both patterns but with the "subcategoriser" pattern [C] more limited than for *in*, and the "semantically coherent nouns" pattern [A] more prominent. A preliminary study of the other most frequent prepositions in English suggests that they can be seen as falling at different points on the spectrum defined by the relative dominance of these two patterns, roughly speaking as follows:

- "Subcategoriser" pattern predominates: *to, by, with, from*
- Neither pattern fully predominates: *in, on, for*
- "Semantically coherent nouns" pattern predominates: *at, of*

Further work is clearly needed here to arrive at a full picture of the relationships and contrasts between these prepositions. A full analysis of all nine lies outside the scope of this study, however – particularly *of* and *to*, which present special problems, as previous analyses have often recognised (see for instance Sinclair, 1991: 81-98). The collocate lists from FLOB for these other prepositions are given in the Appendix (tables A3 to A8). In the next section, findings from the Nepali data will be presented that suggest these patterns are cross-linguistically valid as characterisations of the grammatical category of adposition.

5. Patterns in the Nepali data

The most common postpositions in Nepali are *ko / kā / kī*, *mā*, *le*, *lāī*, and *bāṭa*. *ko / kā / kī* is the genitive postposition. Like *of* in English, it presents certain special problems and will not be discussed further here.

The next most common Nepali postposition is *mā*. This locative marker can be translated variously as *in*, *on*, *to*, *at*, *by*, or *among* (Schmidt *et al*., 1993: 512), and is thus semantically parallel to English *in* and *at*, discussed in the previous section. Its collocates are shown in Table 5. Note that, due to spelling variation in Nepali, some words occur twice in slightly different spellings, for example *rupa* and *rūpa*. This is not seen as problematic: indeed, the fact that *rupa* is such a highly significant collocate even when its frequency is divided across two word-types like this underlines its importance to the analysis.

The patterns observed initially in the discussion of *in* can be noted, to some extent, here as well. Firstly, there are basic, even stereotypical nouns of place and time. These exemplify in Nepali the pattern, noted above for English, of collocation with nouns that cohere with the basic semantics of the adposition. They are marked with [A] in Table 5 in accordance with the notation established previously.

We can also see a pattern of collocates with more abstract meanings, which participate in set phrases in which the locative has a metaphorical sense. Again this is a pattern noted above, and annotated as [B]. The clearest example of this is *rūpa*. For instance, *ko rūpa mā* means "in the capacity of" (and *rūpa mā* following an adjective is also used to mean "in a ADJ way"). Just as the English translation *in* is

metaphorical here, so the original Nepali *mā* is metaphorical. Another example is *ko sambandha mā*, which means "concerning" – literally "in connection of".

No.	Collocate	Translation	Pattern	Frequency	Z-Score
1	rūpa	appearance, form, shape	B	614	46.1
2	ṭhāuṅ	place	A	181	22.4
3	rupa	appearance, form, shape	B	140	22
4	kṣetra	field, region	A	222	21.3
5	thunā	imprisonment		58	21
6	yasa	this		510	18.9
7	sandarbha	connection	B	72	18.7
8	koṭhā	room	A	118	17.5
9	ghara	house	A	271	17.1
10	ādhāra	support	B	138	17
11	viṣaya	topic, matter	B	138	16.5
12	krama	series		78	14.9
13	avasthā	situation, occasion	A	133	14.7
14	deśa	country	A	198	14.6
15	ādhārita	based	B, C	53	14.4
16	bhāga	portion, share, fate, luck		99	14.1
17	ṭhāṅu	place	A	42	13.4
18	pariṇata	change	C	32	13.3
19	sambandha	connection	B	111	13.0
20	khaṇḍa	part, portion, section		58	12.9

Table 5: Collocations of *mā*.

However, some of these abstractions overlap with collocates for which *mā* may be seen to be functioning as a subcategoriser. Consider *ādhāra* "support" and the cognate form *ādhārita* "based". A survey of the relevant concordances shows that these are significant collocates because of their occurrence in the set phrases *ko ādhāra mā* ("on the basis of") and *mā ādhārita* ("based on"). The first of these is an example of the metaphorical locative, as discussed above. The same semantic association seems to be present in *mā ādhārita*, but in this case *mā* is *also* acting as a subcategoriser for *ādhārita*. This is an important point for two reasons. Firstly, it shows that the patterns, which were cleanly separated in the English data, are not *necessarily* non-overlapping. Secondly, it is some evidence for the "subcategoriser" pattern among the collocates of this Nepali postposition, although this pattern appears to be less prominent than it is for either of the English translation-equivalent prepositions discussed in the previous section (*in, at*).

In fact, with only the evidence from *mā*, it would not be possible to argue for the existence of the "subcategoriser" pattern in the collocations of Nepali postpositions. However, some additional examples of that pattern can be seen around another postposition, *bāṭa*, as shown in Table 6. This postposition may be translated as "after (time)" or "by (means)" but usually means "from". Interestingly, the pattern of

semantically compatible nouns does not appear around *bāṭa* (except, possibly, in the case of *mādhyama*).

No.	Collocate	Translation	Pattern	Frequency	Z-Score
1	kasūradāra	offender		23	29.2
2	mādhyama	middle		36	25.3
3	mukta	free, salvation	C	24	21.3
4	niskie	go out, come out[11]	C	14	20.8
5	jūvā	gambling	n/a	4	19.7
6	bañcita	deprived	C	9	19.5
7	bacna	save, protect	C	9	18.6
8	prāpta	received, obtained	C	53	17
9	asat	not true	n/a	4	16
10	tarpha	to, towards		41	15.8
11	janmie	be born, arise	C	10	15.1
12	ragata	blood		20	14.8
13	meśīnarī	machinery	n/a	2	13.9
14	svasnīmānisa	wife	n/a	2	13.9
15	bacāuṅcha	save, protect	n/a (C)	4	13.8
16	āsu	tear	n/a	4	13.8
17	sodhanī	request	n/a	4	13.8
18	chuṭakārā	free	n/a (C)	4	13.8
19	yasa	this		111	13.3
20	niskane	go out, come out	C	9	13.1

Table 6: Collocations of *bāṭa*.

Several of the collocates seen here have very low absolute frequencies (two or four instances); at this level of frequency statistical collocation cannot safely be relied on and it is thus unsafe to utilize these collocates as evidence for any pattern (therefore, they are marked as [n/a] above). This suggests we are close to the limits of this method of analysis for a dataset of the size of the Nepali dealt with in this study. The translation equivalent of *with*, discussed in the previous section, is the Nepali postposition *saṅga*. This word occurs less than 500 times in the dataset (by contrast, *mā* occurs 16,279 times and *bāṭa* occurs 2,435 times). So a direct contrast to *with* is not possible. However, there are two other postpositions which are only slightly less frequent than *mā*: these are *le* and *lāī*, and they will be discussed in turn before moving to a more general discussion of this study's findings. These postpositions have no direct translation equivalent in English, as they mark transitive subjects (*le*) and indirect objects/animate direct objects (*lāī*). Their collocations are shown in tables 7 and 8.

[11] Note that *niskie* (like *niskane* lower down the list) is an inflected verb form. The translations given in these tables are of the verb root only.

No.	Collocate	Translation	Pattern	Frequency	Z-Score
1	mai	I	P (A)	697	72.5
2	usa	that (him/her)	P (A)	501	41
3	hunā	be		145	31.7
4	una	that (him/her)	P (A)	388	30.6
5	gardā	do		222	25.4
6	sarakāra	government[12]	P (A)	278	25
7	kasai	whom (int.)	P (A)	159	22.8
8	āphno	oneself's	P (A)	268	19.1
9	āphū	oneself	P (A)	146	17.9
10	uhāň	there (he/she)	P (A)	150	16.8
11	hāmī	we	P (A)	200	16.4
12	bhanyo	say	C?	82	15.9
13	praharī	police	A	108	15.3
14	harū	PLRL	A?	1087	14.9
15	janatā	people	A	182	14.6
16	kasa	whom (int.)	P (A)	51	14.3
17	tina	there (he/she)	P (A)	116	14.2
18	jasa	whom (rel.)	P (A)	110	13.1
19	kuntā	woman	A	62	12.5
20	bhanin	say	C?	40	12.0

Table 7: Collocations of *le*.

The most notable collocational feature of these postpositions is their strong link with various pronouns. These have been marked with [P] in tables 7 and 8. However, it will be argued here that this is actually a manifestation of the "semantically coherent nouns" pattern (noted as [A]) that was previously observed both for *mā* and for the English prepositions. In this case, the coherence is between the semantic/grammatical roles that the postpositions relate to, and the semantic trait of animacy or humanness.

The transitive subject (marked by *le*) is prototypically an agent, and animacy/humanness is a semantic feature associated with agents (since agency implies the ability to take action intentionally). The indirect object, marked by *lāī*, is prototypically a recipient; recipients are also typically humans or at least animates. And while the semantic role of patient, which is prototypically linked with the grammatical relation of direct object, is not associated with animates more strongly than with inanimates, *lāī* only marks direct objects *when they are animate* (as noted, for instance, by Acharya, 1991: 160; Hutt and Subedi, 1999: 94-95). So we might expect both *le* and *lāī* to be semantically linked to animacy and humanness. To back up this point, note that all the noun collocates of *le*, and two of the four noun collocates of *lāī*, represent human beings. Both *le* and *lāī* also collocate with *harū*, the plural marker; plural number is only explicitly marked on animates in Nepali, so this might, arguably, be considered another link to animacy.

[12] *sarakāra* is also used as the highest-level honorific pronoun in Nepali, reserved for eminent royalty.

No.	Collocate	Translation	Pattern	Frequency	Z-Score
1	ma	I	P (A)	654	52.7
2	āphū	oneself	P (A)	139	25
3	timī	you	P (A)	142	21
4	harū	PLRL	A?	793	19
5	kasai	whom (int.)	P (A)	96	18
6	usa	that (him/her)	P (A)	181	16.3
7	una	that (him/her)	P (A)	170	15.5
8	thāhā	knowledge		85	15.4
9	hāmī	we	P (A)	132	14.8
10	bheṭna	meet	C?	19	14.1
11	janatā	people	A	124	13.9
12	haru	PLRL	A?	188	13
13	dṛṣṭigata	concerning perspective		11	12.6
14	liera	take	C?	68	12.3
15	abhiyukta	culprit		22	11.8
16	yasa	this (him/her)	P (A)	217	11.4
17	upayogasiddha	proven useful	n/a	4	11.1
18	pradāna	gift	C?	30	11
19	giraphtāra	arrest		19	10.8
20	bolāuna	speak	C?	11	10.5

Table 8: Collocations of *lāī*.

This analysis can be extended to incorporate the pronoun collocates by noting that pronouns either represent implicitly animate entities (the first and second person) or else refer back to entities already mentioned. Referents which persist in discourse are frequently animates, especially humans. So the co-occurrence of pronouns with *le* and *lāī* represents the pattern of "semantically coherent" nominals as observed several times in this study. This analysis is supported to an extent by the findings of Genetti and Crain (2003), who report on the basis of a study of spoken Nepali that pronouns are very frequent in Nepali discourse, and that these pronouns are more likely to realise animate referents than inanimate referents.

The other pattern, of the adposition as a subcategoriser for a verb or other collocate (noted by [C]), is possibly attested for *le* and *lāī* in the form of the collocations with forms such as *bhanyo*, *bhanin*, *liera*, and *bolāuna*; however, with relatively little evidence to work from, it is impossible to make a strong claim for this pattern on the basis of the current data.

To repeat the exercise undertaken at the end of the preceding section, the following tentative classification of the Nepali postpositions examined so far according to the relative of the two collocational patterns is proposed:

- "Subcategoriser" pattern predominates: *bāṭa*
- "Semantically coherent nouns" pattern predominates: *mā*, *lē*, *lāī*

Preliminary inspection of the collocations of genitive *ko / kā / kī* suggests that this postposition's collocates are dominated by the "subcategoriser" pattern. However, like English *of* and *to*, this word presents particular problems for the analysis adopted here analysis, and requires more extensive analysis than can be devoted to it here.

6. Discussion

The aim of this paper was to ascertain whether adpositions as a category could be characterised solely in terms of their collocational properties. The answer that has been (tentatively) arrived at is that they can. Two patterns were repeatedly observed in Nepali and English: (i) a pattern of *noun collocates* with meanings *semantically coherent* with the semantics of the adposition; and (ii) a pattern of collocates (mostly *verbs* but also nouns and adjectives) for which the adposition *functions as a subcategoriser*. However, in both languages the category contains an internal gradient, between adpositions characterised predominantly by the former pattern, and adpositions characterised predominantly by the latter pattern.

Depending on one's theoretical perspective, one might argue that these patterns are actually what *constitutes* the grammatical category of "adposition". In other words, if, as Hoey (2005: 154) argues, a category such as "noun" is a shorthand label for a bundle of collocational features such as colligation with *the*, then "adposition" is a shorthand label for words that possess one or the other or both of the collocational patterns that have been observed in the analysis outlined above.

One criticism which might be levelled at this study is that it identified only features of adpositions which are intuitively obvious. For instance, it might be argued that *of course* basic or common nouns of place and time will collocate with *in* or *at* in English, and with *mā* in Nepali: that is what *in*, *at* and *mā* mean. However, there is another way to view this: collocation of those adpositions with such nouns, occurring in the language which an individual speaker hears in their lifetime, is the *cause* of the speaker understanding those adpositions as having those meanings. This then results in the speaker producing, in their turn, utterances with those same collocations in them. To use Hoey's (2005) terminology, it is exactly because *in*, *at* and *mā* are primed for use with nouns of place and time that these adpositions have the meanings that a native speaker intuits them as having.

To put it another way, if collocational analysis like that above leads to conclusions which are also accessible via linguistic intuition, then it is because those linguistic intuitions are *based on* the mind-internal equivalent of a collocational analysis. The value of the method demonstrated here, then, may conceivably be that it lays out explicitly the process by which intuitions about meanings and about grammatical categories are arrived at.

A point of interest here is that adpositions are traditionally seen as markers of *case*. They are thus an example of *dependent marking*, as opposed to *head marking*, in traditional grammatical terms. However, the collocational "subcategoriser" pattern is a link between the adposition and what would traditionally be described as its *head*, not its dependent. It was noted above that a working assumption for this study was that statistical score of a collocation can serve as a proxy measure for the cognitive salience of the association in question. This suggests that the relationship of adpositions and their heads is, for some adpositions, more cognitively salient than the traditional view of them as case markers might imply. But this is a point that would require further investigation.

7. Conclusion: future directions

This study aimed to demonstrate the utility of a "quantitative-distributional" method based on statistical collocation in the investigation of grammatical categories, in this case adpositions. The successful identification of two general patterns that apply across both languages suggests that this aim has been achieved. This relatively brief study cannot however represent the end point of this line of inquiry. It was noted above that *of* and *to* in English, and *ko / kā / kī* in Nepali, present special problems for this kind of analysis. In particular, *ko / kā / kī* and *to* are used in to formation of complex verb structures (the Nepali perfect and the English *to*-infinitive); the impact of this on the type of analysis exemplified in this paper remains to be investigated in detail.[13]

Some noticeable, and surprising, features of the English collocations have not been discussed here. For instance, several of the prepositions collocate very strongly with the definite article, but not all of them (see tables 1 to 4 and the appendix). Given that prepositions occur before noun phrases, and noun phrases often begin with articles, one might expect that all prepositions would show this pattern. Likewise, some prepositions, but not all, collocate with the indefinite article. However, a full investigation of this phenomenon is beyond the scope of the present study.

Some methodological issues remain to be addressed. The choice of the Z-score statistic for the calculation of collocates, over (for example) log likelihood, was dictated by the options available within the Xaira software. The choice of this software was in turn dictated by the Unicode/XML format of the Nepali corpus. Experimentation with other collocation statistics is required, to be certain that the quantitative-distributional method based on statistical collocation is fully rigorous. Work with larger corpora will also be required, to examine the collocations of adpositions less frequent than those discussed above. In particular, as more extensive Nepali corpus data becomes available, it will be possible to revisit the postpositions discussed above, and double-check the findings for *mā*, *bāṭa*, *le* and *lāī*.

In the absence of more extensive analysis of a larger dataset, a larger range of adpositions, and full testing of a range of potentially appropriate statistical measures, the results presented here cannot be a definitive final word on the collocational properties of adpositions in these two languages. However, these preliminary findings are promisingly indicative and suggest that collocation-based methods can be productively applied to the study of grammatical elements.

References

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

Baker, P. (2006) *Using corpora in discourse analysis*. London: Continuum.

Francis, W.N. and H. Kucera. (1964)

Genetti, C. and LD. Crain. (2003) Beyond Preferred Argument Structure: sentences, pronouns, and given referents in Nepali, in J.W. Du Bois, L.E. Kumpff and W.J. Ashby (eds) *Preferred Argument Structure: Grammar as architecture for function*. Amsterdam: John Benjamins.

Greenberg, J. H. (1963) Some universals of grammar with particular reference to the

[13] Some initial comment on *ko / kā / kī* is given in Hardie (forthcoming).

order of meaningful elements, in: J.H. Greenberg (ed.), *Universals of language*, pp. 73–113. Cambridge, Mass.: MIT Press.

Hardie, A. (2004) *The computational analysis of morphosyntactic categories in Urdu*. PhD thesis, University of Lancaster.

Hardie, A. (2005) Automated part-of-speech analysis of Urdu: conceptual and technical issues, in Y. Yadava, G. Bhattarai, R.R. Lohani, B. Prasain and K. Parajuli (eds) *Contemporary issues in Nepalese linguistics*. Kathmandu: LSN.

Hardie, A. (forthcoming). A collocation-based approach to Nepali postpositions.

Hoey, M. (2005) *Lexical Priming*. Routledge.

Hundt, M., A. Sand, and R. Siemund (1998). *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Available online at http://khnt.hit.uib.no/icame/manuals/flob/index.htm .Accessed 30 June 2007.

Hutt, M and Subedi, A (1999) *Nepali*. London: Hodder.

Kennedy, G. (1991) *Between* and *through*: the company they keep and the functions they serve, in K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*, pp. 95–110. London: Longman.

Schmidt, R.L., B.M. Dahal, K.B. Pradham, and G. Vajracharya (eds). (1993) *A practical dictionary of Modern Nepali*. Delhi: Ratna Sagar.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: OUP.

Tomasello, M. (2003) *Constructing a language: a usage-based theory of language acquisition.* Harvard University Press.

Appendix: additional English data

This appendix includes collocation tables that have been alluded to or mentioned briefly, but not extensively discussed, in the foregoing paper.

	BNC Sampler (written)			**BNC Sampler (spoken)**		
No.	Collocate	Frequency	Score	Collocate	Frequency	Score
1	deal [C]	80	38.3	dealing [C]	55	55.9
2	associated [C]	63	37.9	deal [C]	75	49.9
3	dealt [C]	34	35.9	dealt [C]	33	45
4	compared [C]	51	34.7	associated [C]	20	31.7
5	dealing [C]	34	27.8	agree [C]	47	31.1
6	cope [C]	24	26.7	cope [C]	20	28.8
7	filled [C]	34	24.7	happy [C?]	40	26.4
8	contact [C]	37	24.5	start [C]	64	21.5
9	a	1028	24	deals [C]	12	21.1
10	coupled [C]	15	23.7	the	1067	20.6
11	connected [C]	23	21.2	wrong [C]	47	20.2
12	faced [C]	22	20	comply [C]	9	19.7
13	familiar [C]	22	17.6	up	177	19
14	comply [C]	9	17.1	fed up [C]	11	18.8
15	links [C]	21	16.7	wi	15	18.4
16	interfere [C]	7	15.6	coping [C]	6	18
17	teamed [C]	6	15.5	conformance [C]	5	17.8
18	coincide [C]	8	15.1	problems	31	16.3
19	concerned [C]	27	14.9	disagree	5	16.2
20	charged [C]	19	14.7	a	546	16.2

Table A1: Collocations of with in the BNC Sampler.

	BNC Sampler (written)			**BNC Sampler (spoken)**		
No.	Collocate	Frequency	Score	Collocate	Frequency	Score
1	looked [C]	174	60.1	look [C]	511	112.8
2	end [A]	127	39	moment [A]	220	110.8
3	look [C]	118	38.4	looking [C]	186	74.9
4	time [A]	214	38	end [A]	155	56.3
5	aimed [C]	29	33.6	looked [C]	74	43.7
6	looking [C]	72	30.4	the	1466	43.4
7	home [A]	97	30.3	time [A]	233	41
8	beginning [A]	45	28.4	stage [A]	44	36.9
9	stage [A]	55	25.5	home [A]	90	34.6
10	same [A]	93	24.4	school [A]	79	33.2
11	westminster [A?]	30	23.9	beginning [A]	33	31.9
12	glanced [C]	16	23.4	bottom [A]	38	27.2
13	times [A]	57	23.1	o'clock [A]	36	23.1
14	educated [C]	17	23	lunchtime [A]	11	21.9
15	the	1922	22.4	coll	4	19.7
16	staring [C]	13	21.9	top [A]	36	19.1
17	stared [C]	11	21.6	night [A]	54	18.4
18	portman	9	21	barbican [A?]	3	18.2
19	moment [A]	33	20.9	midday [A]	5	17.3
20	arrived [C]	31	20.2	mis	2	17.2

Table A2: Collocations of *at* in the BNC Sampler.

	FLOB		
No.	Collocate	Frequency	Score
1	the	18786	139.2
2	number	390	46.4
3	part	349	39.3
4	kind	241	37.7
5	a	4316	35.1
6	one	843	30.1
7	sort	145	29.8
8	end	261	28
9	thousands	73	25.2
10	some	526	24
11	range	148	23.1
12	importance	91	22.1
13	lot	101	22
14	nature	130	21.5
15	use	225	21.4
16	lack	77	21
17	variety	76	20.7
18	couple	98	20.6
19	rest	110	19.9
20	types	69	19.8

Table A3: Collocations of *of* in FLOB

	FLOB		
No.	Collocate	Frequency	Score
1	be	2815	91.9
2	able	286	50.5
3	go	391	42.9
4	make	404	39.7
5	want	288	39.4
6	going	289	35.6
7	trying	143	34.4
8	get	322	34.2
9	in order	112	31.5
10	have	1040	31.3
11	likely	161	29.9
12	keep	165	28.8
13	him	557	28.5
14	give	203	27.9
15	take	260	27.7
16	tried	118	27.6
17	wanted	140	27.3
18	bring	125	27
19	try	117	26.9
20	come	232	26.4

Table A4: Collocations of *to* in FLOB

	FLOB		
No.	Collocate	Frequency	Score
1	responsible	71	39.1
2	reasons	75	36.8
3	reason	68	25.5
4	waiting	52	25.4
5	purposes	32	23.9
6	min	23	22.7
7	moment	64	21.6
8	looking	63	20.7
9	search	24	20.3
10	years	134	18.5
11	a	1162	17.9
12	suitable	26	17.9
13	searching	17	17.5
14	responsibility	32	17
15	plans	41	16.4
16	sake	17	15.7
17	scope	20	15.5
18	demand	27	15.1
19	quest	11	14.5
20	wait	26	14.5

Table A5: Collocations of for in FLOB

	FLOB		
No.	Collocate	Frequency	Score
1	based	122	57.3
2	the	2781	32.9
3	dependent	30	29.4
4	depend	30	28.7
5	rely	22	27.2
6	earth	43	26.7
7	basis	47	25.5
8	depends	25	25.4
9	emphasis	31	24.4
10	saturday	30	23.5
11	tuesday	18	21.3
12	occasions	26	21.2
13	depended	14	20.8
14	grounds	27	20.3
15	occasion	26	20.2
16	focused	15	19.6
17	sunday	34	19.5
18	friday	21	19.5
19	concentrate	21	19.5
20	went	73	19.3

Table A6: Collocations of *on* in FLOB

	FLOB		
No.	Collocate	Frequency	Score
1	followed	65	36.2
2	supported	26	26.2
3	surrounded	21	25.4
4	dominated	21	24
5	replaced	24	22.6
6	accompanied	19	22.2
7	influenced	17	22.2
8	caused	30	22
9	sponsored	9	20.6
10	affected	21	20.5
11	governed	9	19.6
12	provided	36	19.5
13	supplied	14	18.8
14	inspired	14	18.1
15	backed	14	16.7
16	virtue	16	16.3
17	been	167	15.9
18	owned	12	15.7
19	favoured	12	15.4
20	assisted	8	15.3

Table A7: Collocations of *by* in FLOB

	FLOB		
No.	Collocate	Frequency	Score
1	derived	39	43.3
2	ranging	13	27.5
3	arising	9	23.3
4	the	1578	21.8
5	benefited	10	21
6	far	57	20.4
7	stemmed	6	20.1
8	suffering	16	20
9	dating	11	19.9
10	different	57	19.9
11	borrowed	10	18.6
12	derive	6	18.6
13	removed	18	18.4
14	differs	5	18.3
15	flowed	5	16.7
16	treblinka	2	16.5
17	ravensbrück	2	16.5
18	auschwitz-birkenau-monowitz	2	16.5
19	belzec	2	16.5
20	chelmno	2	16.5

Table A8: Collocations of by in FLOB