

Collocational patterning in cross-linguistic perspective: adpositions in English, Nepali, and Russian

Andrew Hardie and Olga Mudraya
Lancaster University

Abstract

This paper presents a contrastive analysis of adpositions in English, Nepali and Russian corpora. Two sets of highly frequent adpositions, those with broadly locative and broadly ablative meaning, are contrasted. The ‘quantitative-distributional’ analysis is based on identifying patterns across the most statistically significant collocations of the words in question; it is undertaken using 1 million word comparable multi-genre corpora of each language. The results suggest that, while in all three languages the adpositions are characterised by two collocational patterns (one of subcategorisation and one of semantic congruence), the former pattern is substantially more prominent in English than either Russian or Nepali.

Keywords: Collocation, adpositions, English, Nepali, Russian

1 Introduction: A quantitative-distributional approach to adpositions

The premise of this paper is that collocational patterns observed across words within a grammatical category can be used as a means of identifying the distributional characteristics of that category. This methodology, dubbed “quantitative-distributional”, is based on statistical measures of collocation in text corpora. First, tables of highly significant collocates for a number of given search nodes – in this study, the twenty most significant collocates in a span of two words left and two right of each node, based on the Z-score statistic – are assembled. Then, an analysis is conducted into grammatical and/or semantic patterns evident across the collocation tables. These patterns are used to characterise the commonalities across the search nodes, especially in terms of grammatical category membership and internal gradience of categories. Theoretically, this corpus-based methodology is compatible with any of a diverse range of positions, from Hoey’s (2005) theory of Lexical Priming to forms of Construction Grammar such as that presented by Croft (2001).

This method has previously been applied to an analysis of adpositions in Nepali (Hardie 2008) and to a basic-level comparison of English and Nepali adpositions (Hardie 2007). This earlier work argues that the adpositions as a grammatical category are primarily characterised

by two distributional patterns. Firstly, the collocates of adpositions frequently include typical (or stereotypical) nouns of place and time (including terms such as *time* and *place* but also proper nouns of places: countries, cities and so on). Furthermore, the collocates of adpositions typically include nouns with which the adposition forms an idiomatic phrase within which (critically) the adposition has a metaphorical meaning. For example, *fact*, *case* and *context* are all instances of this pattern which occur with English *in*. These two patterns are different instantiations of the same phenomenon, namely collocation of adpositions with *semantically coherent nouns* – where the semantic coherence may be with the literal, concrete meaning of the adposition, or with an abstract, metaphorical sense.

The other noticeable pattern is collocational links between adpositions and lexical items for which the adposition functions as a *subcategoriser*. This term is used to refer to phrasal patterns where the adposition 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 often verbs (for example, English *interested in X*, *involved in X*), but sometimes nouns (*interest in X*, *differences in X*) or adjectives.

In the present paper,¹ this contrastive analysis is extended to Russian prepositions, throwing new light on some aspects of the nature of adpositions as a category.

2 Data

Three broadly comparable corpora have been used in this study to represent the three languages being contrasted. Each corpus is approximately one million tokens in extent. Although much larger datasets exist for all three languages, the advantage of the small datasets used here is that all have been carefully designed to represent a wide range of genres across a detailed sampling frame. FLOB (Hundt et al. 1998) follows the fifteen-genre sampling frame established by the Brown Corpus, as does the ‘Core Sample’ section of the Nepali National Corpus (Yadava et al. 2008). Although it does not follow the same sampling frame, the Uppsala Russian Corpus² (Lönngren 1993) is also balanced across genres, consisting of 50% informative writing (from a range of subject areas) and 50% prose fiction. All corpora were analysed using the same software: CQPweb, a web-based front-end to the IMS Corpus Workbench³. Given the size of the corpora, it was necessary to focus on the most

¹ This research was undertaken as part of the CORGRAM project, a corpus-based investigation of the grammatical categories of three geographically and typologically distinct Indo-European languages. This project was funded by the UK Arts and Humanities Research Council.

² See also <http://www.slaviska.uu.se/korpus.htm>

³ See <http://cwb.sourceforge.net>

frequent adpositions. There are nine very frequent⁴ adpositions in English (*of, in, to, for, on, with, by, at* and *from*), six in Nepali⁵ (genitive *ko / kā / kī*, locative *mā*, ergative-instrumental *le*, accusative-dative *lāī*, plural-collective⁶ *harū*, and *bāṭa* ‘from’) and six in Russian (*v / vo* ‘in, at, on, into, to’, *na* ‘on, in, at, to, for’, *s / so* ‘with, from, and’, *k* ‘to, towards, into’, *po* ‘on, along, over’, and *iz* ‘from, out, of’). We will focus here on collocations of the following, roughly semantically equivalent, groups of adpositions:

- *Broadly locative*: English *in, at, and on*; Nepali *mā*; Russian *v / vo* and *na* (tables 1 to 4)
- *Broadly ablative*: English *from*; Nepali *bāṭa*; Russian *s / so* (tables 5 to 7).

3 Analysis

In the following tables, the collocational patterns discussed above are annotated as follows:

[A] the collocate is a noun whose semantics are congruent with those of the preposition (nouns of place and time).

[B] the collocate is a noun which, with the preposition, forms a phrase in which the preposition has metaphorical meaning (metaphorical semantic congruence).

[C] the collocate is a word for which the preposition functions as a *subcategoriser*.

	at			in			on		
No.	Collocate	Freqy	Score	Collocate	Freq	Score	Collocate	Freqy	Score
1	stared [C]	45	41.6	the	7008	40.5	based [C]	122	57.3
2	look [C]	118	41.5	fact [B]	193	31.5	the	2781	32.9
3	looked [C]	106	39.2	case [B]	194	28.6	dependent [C]	30	29.4
4	aimed [C]	30	36.7	britain [A]	119	21	depend [C]	30	28.7
5	glanced [C]	32	35.2	cases [B]	87	20.6	rely [C]	22	27.2
6	time [A]	184	33	ways	82	20.3	earth [A]	43	26.7
7	end [A]	96	31.3	early	123	20.3	basis [C]	47	25.5
8	home [A]	91	27.2	interested [C]	57	20	depends [C]	25	25.4
9	the	1900	26.7	england [A]	110	20	emphasis [C]	31	24.4
10	looking [C]	54	25.4	detail [B]	53	19.5	saturday [A]	30	23.5
11	staring [C]	15	25	involved [C]	83	19.5	tuesday [A]	18	21.3
12	same [A]	88	24.5	london [A]	137	18.8	occasions [A]	26	21.2
13	intervals [A]	13	24.1	interest [C]	95	18.7	depended [C]	14	20.8
14	expense	17	22.8	context [B]	55	18.5	grounds [B]	27	20.3
15	school [A]	56	22.6	europe [A]	82	18.4	occasion [A]	26	20.2
16	moment [A]	46	21.9	scotland [A]	53	18	focused [C]	15	19.6
17	beginning [A]	32	20.6	france [A]	60	17.4	sunday [A]	34	19.5
18	gazing [C]	10	19.7	america [A]	64	17.3	friday [A]	21	19.5
19	temperatures	14	19.6	middle	58	16.7	concentrate [C]	21	19.5
20	outset [A]	9	19	differences [C]	55	16.4	went [C]	73	19.3

Table 1. Collocations of English at, in, on

⁴ For current purposes, ‘very frequent’ is defined as ‘occurring more often than 4,000 times per million words’.

⁵ The previously published pilot data on Nepali postpositions (Hardie 2007, 2008) was based on a ~40% subset of the NNC-CS. The current data is based on the full first release version of the NNC-CS and differs in some slight respects from the pilot data.

⁶ The Nepali plural marker behaves in many ways, though not all, like an adposition (see Hardie 2007).

No.	v				vo			
	Collocate	Translation	Freq	Score	Collocate	Translation	Freq	Score
1	godu [A]	year	473	57.066	mnogom	much	63	117.2
2	tom	that	646	49.746	-pyervikh	firstly	56	114.856
3	tchisl'ye [B]	number	228	44.264	vsyakom	any	56	114.856
4	ryezul'taty'e [B]	[as a] result	191	40.013	-vtorikh	secondly	44	101.809
5	tchastnosti [B]	particular	188	39.867	snye [B]	sleep	40	93.594
6	etom	this	540	38.975	mnogikh	many	100	83.93
7	stranye [A]	country	227	37.151	vsyem	everything	97	69.997
8	slutchaye [B]	case	235	37.087	glavye [B]	[at the] head	19	63.608
9	usloviyah [B]	conditions	218	35.896	vryemya [A]	time	150	59.78
10	tsyelom [B]	[as a] whole	159	35.678	dvorye [A]	yard	24	58.829
11	nyem	it, him	259	34.705	dvor [A]	yard	32	57.173
12	kontsy'e [A]	end	173	34.512	vsyekh	all	99	50.288
13	storonu [A]	side	202	34.41	slutchaye [B]	case	55	48.675
14	moskvy'e [A]	Moscow	142	34.278	frantsii [A]	France	19	45.893
15	oblasti [A]	region	231	32.573	vtorom	second	20	45.011
16	proshlom [A]	past	138	31.617	rtu	mouth	14	44.683
17	domye [A]	house	128	29.302	vryemyena [A]	times	26	43.396
18	vidye [B]	kind [as]	139	29.247	vtornik [A]	Tuesday	6	34.782
19	obshtchyem [B]	general	114	29.034	mirye [A]	world	31	34.345
20	khodye [B]	course (of)	110	28.982	vsyeoruzhii [B]	armed (with)	5	34.32

Table 2. Collocations of Russian v / vo

No.	na			
	Collocate	Translation	Freq	Score
1	nyesmotrya	despite	138	48.0
2	osnovye [B]	basis	158	44.5
3	urovnye [B]	level	94	36.5
4	myesty'e [A]	place	111	34.6
5	zyemlye [A]	land, earth	114	31.9
6	vzglyad [B]	view	111	31.8
7	byeryegu [A]	beach	87	30.6
8	etapy'e [A]	stage	59	30
9	ulitsy'e [A]	street	72	28.9
10	stolye [A]	table	52	28.3
11	glyadya [C]	looking (at)	75	27.4
12	dyelye [B]	deed	105	25.7
13	ulitsu [A]	street	52	24.5
14	zyemlyu [A]	Earth, ground	91	24.4
15	zavodye [A]	works, factory	41	24.2
16	pyedpriyat'yakh [A]	plant, enterprise	39	23.9
17	samom [B]	actual	103	23.9
18	protyazhnyii [A]	duration, length	33	23.1
19	posmotryela [C]	looked (at)	39	22.4
20	svyetye [A]	world, light	61	22.2

Table 3. Collocations of Russian na

No.	mā			
	Collocate	Translation	Freq	Score
1	rūpa [B]	appearance, form, shape	1603	78.2
2	ṭhāuñ [A]	place	552	39.8
3	rupa [B]	(as 1)	342	37.5
4	kṣetra [A]	field, region	551	32.9
5	ādhāra [B]	support	371	32.0
6	krama	series	260	29.6
7	avasthā [A]	situation, occasion	404	29.3
8	sandarbhā [B]	connection	175	28.5
9	yasa	this (him/her)	1229	27.2
10	ghara [A]	house (home)	628	26.2
11	sambandhā [B]	connection	323	25.6
12	samaya [A]	time	422	23.3
13	ādhārita [B], [C]	based	110	22.8
14	koṭhā [A]	room	222	22.4
15	bhāga	portion, share, fate	250	21.7
16	mātrā [B]	quantity	140	21.5
17	āpasa	oneself	81	21.3
18	viśaya [B]	topic, matter	278	21.2
19	deśa	country	441	20.3
20	sāla [B]	year	251	20.1

Table 4. Collocations of Nepali mā.

from			
No.	Collocate	Freq	Score
1	derived [C]	39	43.3
2	ranging [C]	13	27.5
3	arising [C]	9	23.3
4	the	1578	21.8
5	benefited [C]	10	21
6	far [A]	57	20.4
7	stemmed [C]	6	20.1
8	suffering [C]	16	20
9	dating [C]	11	19.9
10	different [C]	57	19.9
11	borrowed [C]	10	18.6
12	derive [C]	6	18.6
13	removed [C]	18	18.4
14	differs [C]	5	18.3
15	flowed [C]	5	16.7
16	came [C]	51	16.4
17	deriving [C]	4	16.4
18	detract [C]	4	16.4
19	borrowing [C]	10	15.7
20	escape [C]	16	15.7

Table 5. Collocations of English from

bāṭa				
No.	Collocate	Translation	Frey	Score
1	mukta [C]	free, salvation	69	43.1
2	mādhyaṃa [B]	means, medium	92	37.5
3	tarpha	to, towards	120	28.0
4	prāpta [C]	received, obtained	127	27.1
5	bacna [C]	save, protect	18	24.9
6	bāhira [A]	out	84	23.5
7	bañcita [C]	deprived	15	23.5
8	ṭāḍhai [A]	farther	14	21.3
9	kasūradāra	offender	13	20.9
10	ṭāḍhā [A]	farther	44	20.8
11	niskie [C]	go out, come out	15	19.7
12	mukha	mouth	64	19.5
13	mausūph	majesty	17	19.3
14	yasa	this (him/her)	271	18.8
15	kampyuṭara	computer	23	18.6
16	niskera [C]	going out	14	18.4
17	tyahāñ	that / there	92	17.6
18	niskane [C]	go out, come out	15	17.2
19	prabhāvita [C]	impressed	28	17.1
20	umkana [C]	become free	6	16.3

Table 6. Collocations of Nepali bāṭa.

s					so			
No	Collocate	Translation	Freq	Score	Collocate	Translation	Freq	Score
1	vmyestye [C]	together	312	72.8	mnoi	me	67	104.9
2	pomoshtch'yu [B]	help	200	68.8	vryemyenyem [A]	time	50	94.9
3	nim	him, it	295	60.1	vsyemi	all	59	93.1
4	uchyetom [B]	taking into account	70	42.1	storoni [B]	side, outside	92	85.9
5	ryadom	next (to), a number of	152	40.1	svoimi	one's	46	64.9
6	srvnyenyiu [C]	comparison	67	39.3	storon [B]	sides	24	47.1
7	svyazi [C]	connection	127	38.6	skorost'yu [B]	speed	14	40.9
8	nimi	them	126	37.9	dvora [A]	yard	15	40.9
9	storoni [B]	side	140	36.8	svoim	one's	40	40.2
10	toboi	you	58	34.9	skam'i [B]	bench	6	38.8
11	tochki [B]	point	74	34.5	strakhom	fear	9	35.2
12	naryadu	along (with)	50	34.5	slyedami	traces	5	34.9
13	trudom [B]	difficulty	62	34.2	stonom	moan	5	34.9
14	udovol'stviyem [B]	pleasure	46	33.7	svoistvyennoi [C]	characteristic	5	34.9
15	vami	you	48	30.3	svistom [B]	whistle	7	34.5
16	etim	this	123	29.9	szhit'sya [C]	get used to	2	34.3
17	svyazaniye [C]	connected	38	29.4	vyetlami	willows	2	34.3
18	svyazannikh [C]	connected	42	29.3	stranami [A]	countries	14	34.0
19	sootvyetstvii [C]	accordance	39	28.7	shtokom	rod	4	30.6
20	drugimi	other	65	28.0	svyazistami	signalmen	3	29.7

Table 7. Collocations of Russian s / so

It should be noted that it is not considered necessary within this method for *every* collocate of a given search term to be assigned to one of the identified patterns. Each word possesses unique phraseology, which the collocation statistics frequently reflect. Clearly, there are very few or no similarities at the level of individual collocates between the three languages. This is largely what we should expect given that much phraseology is purely conventional in nature. For instance, a comparison of the collocations in tables 5 to 7 that are analysed as instances of [C] shows little or no overlap. However, a comparison across different languages *is* meaningful at the level of the patterns described in §1.

4 Discussion: contrasts across languages

It is immediately obvious that *both* patterns are observed across all three languages. However, there are some subtle differences between English on the one hand, and Nepali and Russian on the other. The adpositions with broadly locative meaning show this contrast: in English, the subcategorisation pattern [C] is relatively more prominent than it is for its Nepali and Russian equivalents – for *v/vo* and *mā* the semantic congruence pattern [A,B] is entirely predominant, and for *na* it is substantially more prominent. There *are* instances of [C] for the Nepali and Russian locative adpositions, but they are found rarely, or not found, amongst the most significant collocations. The broadly ablative adpositions, by contrast, show the same contrast, but on a different basis. For Nepali *bāṭa* and Russian *s/so*, both the semantic congruence [A,B] and subcategorisation [C] patterns are present, with [C] being perhaps slightly more predominant. But for English *from*, [C] absolutely predominates. So for both locative and ablative adpositions, while both patterns are in evidence to *some* degree in all three languages, in English the subcategorisation pattern [C] is relatively much stronger (this contrast is summarised in table 8).

	<i>English</i>	<i>Russian</i>	<i>Nepali</i>
Broadly locative	Neither pattern dominant <i>at</i> (C 8/20, A,B 9/20) <i>in</i> (C 4/20, A,B 12/20) <i>on</i> (C 11/20, A,B 8/20)	[A,B] dominant <i>na</i> (C 2/20, A,B 17/20) <i>v</i> (C 0/20, A,B 16/20) <i>vo</i> (C 0/20, A,B 11/20)	[A,B] dominant <i>mā</i> (C 1/20, A,B 15/20)
Broadly ablative	[C] dominant <i>from</i> (C 18/20)	Neither pattern dominant <i>s</i> (C 6/20, A,B 6/20) <i>so</i> (C 2/20, A,B 8/20)	Neither pattern dominant <i>bāṭa</i> (C 9/20, A,/B 4/20)

Table 8. Summary of relative weightings of the observed patterns across languages

It is interesting to speculate why this might be. It may stem from differences in the overall

inventory of adpositions in the three languages and the way in which different functions are distributed across adpositions. A wider survey of adpositions may shed light on this. Another possible factor may be differences in other means of indicating subcategorisation relations in these languages, and the ways these means interact with the adpositions in each language. In particular it is notable that the grammatical roles of subject and object are marked explicitly in Nepali and Russian (respectively, by ergative-instrumental and accusative-dative postpositions, and by case inflections), whereas English relies primarily on word order. It is possible that there is a link between English's lack of case marking of the main verbal subcategories on the one hand, and its more extensive use of adpositions as explicit indicators of subcategorisation on the other. In other words, English may make more extensive use of locative and ablative (and possibly other) prepositions to mark subcategorisation because it cannot use core grammatical case marking to accomplish this task. However, prior to a comparative analysis of data for a wider range of adpositions, this explanation must remain speculative.

5 Conclusion

To summarise, the picture that emerges from a quantitative-distributional analysis of locative and ablative adpositions in English, Russian and Nepali is that Nepali and Russian are overall, more similar to one another and distinct from English; but adpositions in all three languages are characterised by the same two patterns of collocational behaviour. It has not been possible in the present study to devote any attention to the particular characteristics of any particular adposition, or indeed any particular language, although there are most definitely unique individual phenomena to be observed in the collocation data, alongside the two overarching patterns discussed above. For example, it is notable that in English the definite article collocates with all the adpositions considered here; however, this collocation tends to be more highly significant for the prepositions where the semantic-coherence pattern is more prominent. Collocation data relating to other English prepositions confirms this trend, which remains to be explained. Another point which has not been explored here is the collocation of adpositions with pronouns, which is noticeable in Nepali and Russian but not English (the same split observed for the main patterns discussed in this paper). Hardie (2008) argues that collocation with pronouns is a significant aspect of the semantic-coherence pattern for the Nepali ergative-instrumental and accusative-dative postpositions; the extent to which this explanation will hold for Russian and English remains to be investigated.

A more sophisticated comparison than was possible here would be not between individual

adpositions, but between the *landscapes* of adpositions existing within each language. Further research currently in hand will extend this analysis to a wider range of both adpositions and datasets in all three languages. In particular, the lack of any spoken data in this analysis is an acknowledged weakness of this study. However, an initial analysis of spoken data in English (Hardie 2007) suggests that, while the actual collocates found around prepositions vary between speech and writing, the general *patterns* that these collocates instantiate are not.

References

Croft, W. (2001) *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Hardie, A. (2007) Collocational properties of adpositions in Nepali and English. In: *Proceedings of the Corpus Linguistics 2007 conference*. Available online at <http://ucrel.lancs.ac.uk/publications/CL2007/>

Hardie, A. (2008) A collocation-based approach to Nepali postpositions. *Corpus Linguistics and Linguistic Theory* 4(1): 19-62.

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

Yadava, Y.P., Hardie, A., Lohani R.R., Regmi B.N., Gurung, S., Gurung, A., McEnery, T., Allwood, J., and Hall, P. (2008). Construction and annotation of a corpus of contemporary Nepali. *Corpora* 3(2): 213-225.

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>

Lönngren, L. (ed.) (1993). *Chastotnyj slovar' sovremennogo russkogo jazyka*. (A Frequency Dictionary of Modern Russian. With a Summary in English.) Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia.