A corpus based analysis of English, Swedish, Polish, and Russian prepositions

Barbara Gawronska[1], Olga Nikolaenkova[1,2] and Björn Erlendsson[1]
[1] School of Humanities and Informatics, University of Skövde, Sweden
[2] Department of Linguistics, University of Athens, Greece

Abstract

In this study, the use of most frequent spatial prepositions in English, Polish, Swedish, and Russian is analyzed. The prepositions and their contexts are extracted from corpora by means of concordance tools. The collostructional strength between the prepositions and the most frequent nouns in the PPs (Gries et al. 2005) is then computed in order to get a more detailed picture of the contexts in which a given preposition is most likely to appear. The results of the investigation are then analysed within the framework of cognitive semantics, especially Croft and Cruse's (2004) taxonomy of construal operations, and Talmy's (2005) classification of spatial images.

Background and aim

Prepositions define relations between objects, or, rather, conceptualizations of objects. In order to define their meaning it is necessary not only to describe the relation the preposition expresses but also the objects involved. Still, a description based on geometrical notions (the dimensionality of the objects) does not cover all aspects of the semantics of prepositions. Croft and Cruse (2005) propose a model that enriches the geometrical descriptions with construals as focus, scale of attention, perspective and viewpoint.

The main difficulty in cross-linguistic description of preposition semantics is due to the fact that cross-language differences in prepositional systems increase as we move from physical senses of prepositions into the metaphoric extensions of their meaning. The meaning chains (Brugman 1981, Gawronska 1993, Taylor 1988) have different shapes in different languages. Another difficulty lies in different degrees of lexicalization and in formulating criterions for regarding prepositions as parts of lexicalized multiword entries. These problems are of central importance for language technology application, especially Machine Translation (MT).

In the present study, we investigate whether the collostructural analysis, proposed by Gries at al. (2005), may be helpful in identification of nouns and noun classes that tend to co-occur with certain prepositions in 4 different languages, and whether the values of the collostructural strength may contribute to a better understanding of similarities and differences among prepositional systems. Another question we are concerned with is whether and how the results may improve the treatment of prepositions in MT.

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece.

Method and results

The frequency of prepositions in English, Swedish, Polish, and Russian was investigated using large corpora:

A corpus of modern English prose (31 234 174 word tokens)
A corpus of modern Swedish prose (43 634 620 word tokens)
A corpus of modern Polish prose (49 478 901 word tokens)
A corpus of modern Russian (25 000 000 word tokens)

The English, Swedish, and Polish corpora were obtained from LexwareLabs AB (www.lexwarelabs.com), and the Russian one – from the Russian Academy of Science. PPs with three most frequent spatial prepositions for each language were subjects for further investigation. The preposition were: *in, at, on* (English), *i, på, till* (Swedish), *w, na, przez* (Polish), and *в, на, через* (Russian).

The collostrucural strength between the prepositions and the nouns was computed according to the formula in Figure 1. The variables are explained in Table 1. Tables 2-3 show the results for the ten most frequent nouns co-ocurring with with the English and Polish prepositions investigated here.

$$CS = -\mathrm{Log}_{10}\left(\frac{\binom{(a+c)}{a} \times \binom{(b+d)}{b}}{\binom{N}{(a+b)}} \right)$$

Figure 1. The formula for calculation of collostructural strength (CS).

Table 1. An example showing the values of the variables in Figure 1

	Construction C	**Other Constructions**	**Row Totals**
Word = bank	a = f(på + bank)	b= f(other P + bank)	a+b
Other Words	c = f(på+other noun)	d =f(other P+other noun)	c+d
Col. Totals	a+c	b+d	N=a+b+c+d

A comparison of the obtained CS-values showed that the following nouns and noun categories displayed either both high frequency value and high (> 0.6) collostructural strength, or high collocational strength in at least three of the four languages: WORLD, EARTH, COUNTRY, NET/WEB/ INTERNET, WAY/ROAD, TIME, TIME PERIOD, TIME BOUNDARY (start/end), AUTHORITIES, TEXT/NEWSPAPERS/LITERATURE SUBJECT/ MATTER, HEAD.

Table 2. English: Ten most frequent noun co-ocurring with in, on and at in P+N and P+Det+N phrases. F= frequency in thousands, CS= collostructural strength.

In **(90 093 occurrences)**			**On** **(20 119 occurrences)**			**At** **(32 262 occurrences)**		
Noun	**F**	**CS**	**Noun**	**F**	**CS**	**Noun**	**F**	**CS**
world	4,13	2,90	deck	5,57	8,25	moment	3,03	3,57
way	3,57	2,53	account	1,63	1,80	length	2,99	2,74
spite	2,51	2,30	board	1,51	1,80	home	2,89	2,74
front	2,27	2,30	earth	1,00	1,80	end	2,34	2,74
fact	2,37	1,86	contrary	1,19	1,80	rate	1,42	1,60
morning	2,49	1,86	subject	1,83	1,51	night	1,52	1,31
midst	1,69	1,15	ground	1,39	1,34	door	1,66	1,14
house	2,08	1,14	side	1,97	1,34	time	0,82	0,06
love	1,75	0,64	foot	0,62	0,01	hand	0,82	0,02
time	1,96	0,58	deck	5,57	8,25	work	0,73	0,01

Table 3. Polish: Ten most frequent nouns co-ocurring with w, na, and przez

w **388 056 occurrences**			**na** **154 907 occurrences**			**przez** **41 067 occurrences**		
noun	**F**	**CS**	**noun**	**F**	**cs**	**noun**	**F**	**CS**
case	24,5	3,74	ground	23,0	14,20	sms	6,44	7,16
service	23,0	3,38	sake	10,0	3,92	authorities	2,40	1,63
end	17,2	3,31	territory	8,07	3,39	person	2,34	0,99
august	8,82	1,54	subject	7,55	3,41	moment	3,06	0,93
content	6,68	1,49	example	4,18	2,38	period	1,05	0,45
matter	15,7	1,32	earth	7,30	1,70	author	0,81	0,43
Google	6,91	1,15	side	8,21	1,39	people	1,57	0,42
case	9,28	1,07	conclusion	5,70	1,38	head	0,41	0,31
newspapers	6,93	1,05	principle	4,38	1,22	moment	0,39	0,06
goal	17,8	0,80	world	5,27	1,07	agency	0,66	0,03

Conclusions

Our results confirm Gries' et al.(2005) claim that the collocstructural strength value outperforms raw frequency data in corpus-based analysis. For example, although the Polish and Swedish nouns TIME are not among the 10 most frequent nouns after *på/na*, the CS-values between *på* and TIME and *na* and TIME are higher than the values of the 10th most frequent nouns co-occurring with these prepositions, which is intuitively correct. Nevertheless, a high CS-value cannot be used for automatic selection of translation equivalents in Machine Translation without further refinement. Both

Swedish *i* and English *in* have high CS-values in connection to the noun MORNING, but the Swedish phrase *i morgon* is equivalent to *tomorrow*. Collocations with the lexeme TIME should be coded in the lexicon as patterns like:

P + TIME (= the noun "time")+ [viewpoint]

The same is true about the collocations with the categories TIME PERIOD, TIME BOUNDARY.

Furthermore, our analysis reveals certain different conceptualizations of common-experience concepts:

WORLD is a ‘surface’ in Polish and a ‘container’ in Swedish, Russian, and English

WAY/ROAD is 2-dimensional both in the spatial and the metaphorical sense in Swedish and English (vara på väg, be on the way); however, in English it is 2-dimensional if the travellers viewpoint is preserved, and 3-dimensional from an outside perspective (in this way). In Polish, the situation is opposite: WAY is 3-dimensional from the traveller’s point of view (jestem w drodze – ‘I am on the way’), and 2-dimensional otherwise.

Low collostrucurtal values (<0.5) seem to indicate either valence-boundedness of the type V + P or A + P, or a particular syntactic construction on sentence level (e.g. passive; cf. the results for the Polish przez, which is used as agent marker in passive). This hypothesis has to be tested in further research.

References

Brugman, Claudia M. 1981, Story of OVER, Master’s thesis, university of California, Berkeley. Trier: LAUT 1983.

Croft W. and Cruse D. A. 2004: Cognitive linguistics. Cambridge: Cambridge University Press.

Gawronska, B. 1993. Entailment in Logic and in the Lexicon. In: Martin-Vide, C. Current Issues in Mathematical Linguistics. Amsterdam: Elsevier, pp. 239-248.

Gries, S.Th, Hampe, B., and Schönefeld, D. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. Cognitive Linguistics 16-4, pp. 635-676.

Talmy, L. 2005, The fundamental system of spatial schemas in language, In Hampe, B. (ed.), From perception to Meaning, Berlin/New York: Mouton de Gruyter, pp 199-234.

Taylor, J. R. 1988, Contrasting Prepositional Categories: English and Italian, In Rudzka-Ostyn, B. (ed.), Topics in Cognitive Linguistics, Amsterdam/Philadelphia, John Benjamins Publishing Company, pp 299-326.