Second ACL Workshop on Multiword Expressions: Integrating Processing, July 2004, pp. 17-23

What is at stake: a case study of Russian expressions starting with a preposition

Serge Sharoff
Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds, Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk

Abstract

The paper describes an experiment in detecting a specific type of multiword expressions in Russian, namely expressions starting with a preposition. This covers not only prepositional phrases proper, but also fixed syntactic constructions like *v techenie* ('in the course of'). First, we collect lists of such constructions in a corpus of 50 mln words using a simple mechanism that combines statistical methods with knowledge about the structure of Russian prepositional phrases. Then we analyse the results of this data collection and estimate the efficiency of the collected list for the resolution of morphosyntactic and semantic ambiguity in a corpus.

1 Introduction

Computational research on multiword expressions (MWEs) has mostly addressed the topic for English (Sag et al., 2001). Some research has dealt with other languages, such as French (Michiels and Dufour, 1998) or Chinese (Zhang et al., 2000), but there has been no computationally tractable research on the topic for Russian. What is more, the study of MWEs in English has been mostly devoted to the description of nominal groups or light verbs, e.g. (Calzolari et al., 2002), (Sag et al., 2001), while constructions starting with a preposition, such as *in line, at large*, have not been the focus of attention.

Even though the tradition of studying Russian idiomatic expressions resulted in many descriptions of Russian idioms and phraseological dictionaries, like (Dobrovol'skij, 2000) or (Fedorov, 1995), the studies and dictionaries often concentrate on non-decomposable colourful expressions of the 'kick-the-bucket' type, such as *byt' bez carja v golove* ('to have a screw loose', lit. 'to be without a tsar in one's head') and pay no attention to the very notion of their frequency. However, many expressions of this sort are relatively rare in modern language. For example, there is no single instance of *bez carja v golove* in the corpus we used. At the same time, existing Russian dictionaries of idioms often miss more frequent constructions, which are important both for translation studies and for the development of NLP applications. The task of the current study is defined by the ongoing development of the Russian Reference Corpus (Sharoff, 2004), a general-purpose corpus of Russian that is comparable to the British National Corpus (BNC) in its size and coverage. The goal of the study was to identify the list of statistically important MWEs in the corpus and to use them to reduce the ambiguity in corpus analysis.

Existing research on the detection of MWEs can be positioned between two extremes: linguistic and statistical. The former approaches assume syntactic parsing of source texts (sometimes shallow, sometimes deep to identify the semantic roles of MWE components) and the ability to get information from a thesaurus. Detection results can be further improved by deep semantic analysis of

source texts (Piao et al., 2003). When we apply such techniques to a Russian corpus of the size of the BNC, this means that we need accurate and robust parsing tools, which do not exist for Russian. Also, no electronic thesaurus, such as WordNet (Miller, 1990), is available for Russian. Purely statistical approaches treat multiword expressions as a bag of words and pay no attention to the possibility of variation in the inventory and order of MWE components. Given that the word order in Russian (and other Slavonic languages) is relatively free and a typical word (i.e. lemma) has many forms (typically from 9 for nouns to 50 for verbs), the sequences of exact N-grams are much less frequent than in English, thus rendering purely statistical approaches useless.

This paper discusses a hybrid approach to the identification of a specific type of MWEs in Russian, namely constructions starting with prepositional phrases with the emphasis on those that are frequent in the corpus. The study is also aimed at a specific task, namely the disambiguation of their morphological properties and syntactic functions in a corpus. The approach assumes the development of a list of MWEs supported by computational tools, including the calculation of standard statistical measures and shallow parsing of prepositional phrases. In addition, the scope of the study is further distinguished by the goal of extracting MWEs from the core lexicon on the basis of a general-purpose corpus, while many other MWE detection studies concerned the extraction of technical terms specific to a particular domain.

2 The analysis of the structure of Russian MWEs

First, a few words on the linguistic features of MWEs in Russian in general and of prepositional phrases in particular. Russian is an inflecting language in which a word inflects for a set of morphological categories and shows a specific combination of these categories in its ending. For instance, a noun in Russian has a fixed gender and inflects for 6 to 9 cases and for the number (singular or plural, with relics of the dual, which is relevant for some words). Similarly, an adjective inflects for six cases, two numbers and three genders and agrees with the noun that is the head of the nominal group in the values of these three categories. This means that an approach that treats MWEs as 'words with spaces inside' is not always suitable for English, and cannot work for Russian. There is a certain variation in the number of forms in an MWE like *rara avis* in English, because *rarae aves* and *rara avises* are both possible according to (OED, 1989), even though they are extremely rare (neither is used in the BNC and Internet searches mostly point to entries in dictionaries), but at least it is feasible to list the two extra forms separately. At the same the Russian expression *belaja vorona* (corresponding to *rara avis*, lit. 'white crow') exists in 10 different forms (see examples in Table 1, the endings are underlined) and the variability of forms applies to *any* nominal group. The situation is even more complicated in the case of MWEs including verbs, given that in addition to several proper verbal forms, a Russian verb can exist in the form of up to four participles, each of which is inflected as an adjective with its own set of forms.

(1)	*beloj vorony*	genitive, singular
(2)	*beloj vorone*	dative, singular
(3)	*belye vorony*	nominative, plural

Table 1: Examples of the ambiguity of forms

At the same time the large number of forms does not mean that each form can be mapped to a lemma and a set of morphological categories without any ambiguity, because the number of endings is much smaller than the number of possible combinations of features. As lines (1) and (2) in Table 1 suggest, the genitive and dative forms of singular feminine adjectives coincide, as well as the genitive singular and nominative plural forms of the noun *vorona*, see lines (1) and (3).[1]

If we consider prepositional phrases, the amount of ambiguity is much smaller, because prepositions govern the case of a nominal group that follows them and do not themselves inflect.[2] However, PPs still exhibit the general problem of

[1] See (Hajič and Hladká, 1998) for a general overview of problems with the identification of a tagset and the resolution of the ambiguity in Slavonic languages. Their description is about Czech, but it can be applied to Russian as well.

[2] The terminology that distinguishes groups and phrases, e.g. nominal groups vs. prepositional phrases, follows (Halliday, 1985).

ambiguity in lemma selection. For instance, the word form *tem* is ambiguous between the genitive plural form of the noun *tema* (topic) and the instrumental singular masculine form of the demonstrative pronoun *tot* (that). What is more, the prepositional phrase *s tem* from the purely syntactic viewpoint can be interpreted in both ways, because the preposition *s* can govern either the genitive or the instrumental case. At the same time the word *tem* as the component of *s tem chtoby* (in order to, lit. 'with that to') shows no ambiguity in its part of speech. More frequently ambiguity concerns the selection of a lemma or morphological properties for the collocate. For instance, the second word in the expression *s bol'shim zapasom* (with a huge margin, lit. 'with large storage') can be analysed as either of two adjectives *bol'shoj* (large) or *ból'shij* (larger). Similarly, the last word in the expression *do six por* (until now, lit. 'before this time') can be analysed as either of two nouns *porá* (time, season) or *póra* (pore). However, the expressions as a whole are not ambiguous and have specific meanings.

The second problem with prepositional phrases concerns their syntactic function, in particular the notorious PP attachment problem. Even though MWEs consisting of a preposition followed by a nominal group are often identical in their syntactic structure to fully compositional prepositional phrases, they do not carry the same syntactic function as the latter. Such MWEs function in the syntactic structure of the clause as a single unit with a clearly defined meaning that cannot be decomposed into the meaning of their components. In the end, it is better to treat them as adverbs, e.g. *v chastnosti* (in particular), *pod kljuch* (turnkey, lit. 'under key'), or as prepositions in their own right, e.g. *v techenie* ('in the course of'). Multiword expressions starting with a preposition in English have similar structure, but the difference with Russian is that there is no change in the structure of the prepositional group, unlike some English MWEs, e.g. *in line, at large*, which do not have a determiner. Thus, we cannot use the difference in the PP structure as an indicator of an MWE.

The fact that MWEs are not fully compositional means that the meanings of their constituent words change resulting a specific idiomatic meaning of the whole contstruction. In this case we cannot accept the general assumption of one sense per discourse (Gale et al., 1992), because words such as *line, large* in English or *kljuch* in Russian can function in the same discourse in a totally different sense. However, the assumption of one sense per collocation can hold, because an MWE with a prepositional phrase typically has one and the same meaning: even though *line, large* or *techenie* are ambiguous, *in line, at large* and *pod kljuch*, *v techenie* have their specific meanings.

3 Methodology

The study starts with the selection of the list of the most frequent prepositions to account for a large number of potential collocations. Information on the frequency of prepositions (Table 2) is taken from the pilot version of the Russian Reference Corpus, which currently consists of about 55 million words (Table 2 lists the relative frequency of prepositions in terms of the number of their instances per million words, ipm).

Then for each preposition we extract its most frequent collocations in the same corpus and weight them according to the pointwise mutual information score (MI score) and Student's *t* test (T score). Two types of collocates are extracted: all lexical items occurring immediately on the right of a preposition and the longest possible nominal groups defined as the sequence of adjectives and nouns with the condition that nouns after the first one are in the genitive case. This simple pattern captures the majority of Russian nominal groups, except those with elaborations of other clauses or other prepositional phrases embedded inside them. Anyway, because of their nature they do not belong to the class of fixed expressions under study. The MI score foregrounds collocations in which the second component rarely (almost never) occurs outside of the expression in question, whereas the T score foregrounds the most stable collocations on the basis of their frequency.

For every preposition and the list of its most significant collocates we select MWEs on the basis of the lack of compositionality, namely that there is a specific function performed by the expression and this function cannot be automatically derived from the meaning of the words compris-

ing the candidate MWE. The criterion cannot be defined precisely, but in many cases it is immediately obvious that the candidate MWE is or is not fully compositional. For instance, the expression *bez vsjakoj svjazi* ('for no apparent reason', lit. 'without any connection') is sufficiently frequent (38 instances) and the last element has a lexical ambiguity *svjaz'*: connection (either physical or logical) or communication. When the MWE is used in texts, it has a specific function, namely someone's discourse is evaluated as lacking a continuity. Thus, *bez vsjakoj svjazi* is treated as an MWE. On the other hand, the expression *v Rossii* (in Russia) is much more frequent and statistically significant (14557 instances, its T score is 104.21), but the set of locations constitutes an open list, in which other members may be also frequent, e.g. *v SSHA* (in the USA, 4739 instances), *v Evrope* (in Europe, 2752), *v Parizhe* (in Paris, 2087), *v Kitae* (in China, 1055), and the expressions are fully compositional. None of them are considered to be MWEs. At the same time, an expression with a very similar structure: *v storone* ([to keep] aloof, lit. 'in side', 9690 instances, its T-score is 83.95) is considered to be an MWE, because it is not compositional. The vast majority of uses of this expression do not refer to a physical location, but to the fact that a person does not take part in a joint activity.

Also, because of the idiomaticity of the meaning of an MWE, it functions as a whole in the syntactic structure of the clause, most typically as an adjunct, and is translated to other languages in a specific way not necessarily related to prepositional phrases. The possibility of its translation into English without the use of a prepositional phrase is another reason for treating the expression to be a potential MWE.

Finally, an easy test for detecting an MWE concerns the "penetrability" of the expression, i.e. the possibility to insert another word, most typically an adjective or a determiner, into the candidate MWE. If any insertion is unlikely or the meaning of components is redefined as the result of insertion, then the expression in question is an MWE. For instance, even though the MWE *v storone* can be modified as *v drugoj/levoj/protivopolozhnoj storone* (on the other/left/opposite side), the resulting expressions refer to physical locations and not to the idiomatic meaning of the MWE *v storone*. Thus, they are not considered as MWEs but the possibility of insertion here does not violate the penetrability of the MWE in question.

Word	Gloss	Frq (ipm)	Scored	Selected
v/vo	in	27966	703	198
na	on	16513	198	117
s/so	with	11131	734	64
po	over	5816	124	56
k	to	5468	157	20
u	at	4956	203	6
iz	out of	4816	194	6
za	behind	4711	115	34
ot	from	3540	118	13
o	about	2956	357	5
dlja	for	2302	164	13
do	before	1978	477	40
pod	under	1467	139	95
pri	by	1163	140	2
bez	without	1097	459	42
mezhdu	between	502	102	9
Total			4384	720

Table 2: The list of prepositions and the number of their patterns

4 Results

The automatic procedure detected 4384 candidate expressions, out of which we selected 720 MWEs. The summary of prepositions and the number of their patterns identified in the study is given in Table 2. It was expected that more frequent prepositions participate in a larger number of MWEs. However, the situation is more complex. Some prepositions like *u* or *iz* occur almost exclusively in fully compositional patterns, for example, expressing location: *u okna, morja* (by the window, by the sea), or possession: *u menja, u Ivana* (I have, Ivan has). Other prepositions that are less frequent regularly produce non-compositional patterns, e.g. *pod rukami* ('at hand', which expresses the specific meaning of availability, not literally 'under hands'), *pod konec* ('at the end').

The results retained in the database include well-formed prepositional phrases that function as

proper idioms, as well as syntactic constructions that can take a noun or another nominal group on their right, such as *v techenie* (‘in the course of’), which is a PP in its own, or an incomplete combination of a preposition and an adjective such as *dlja puschij* (‘for greater’). The latter is a part of an open list of well-formed PPs, as in *dlja puschej vazhnosti*, (‘for greater importance’), *soxrannosti* (safety), *ostrastki* (frightening), but the word *puschij* in itself occurs only in this construction. In other cases, the ‘noun’ from the nominal group does not even exist in the contemporary language, like in *bez umolku* ([to talk] without a pause), so the expression cannot be analysed correctly without knowing that it is an MWE.

The resulting list also includes multiword expressions with a slightly different structure, in cases where an MWE naturally extends to the left of the prepositionto form a larger pattern. One example is *sudja po vsemu* (‘to all appearances’, lit. ‘judging over all’), which is an extension of a prepositional phrase *po vsemu*, as it gives the only suitable pattern by far with 1626 instances in the corpus, with the next most frequent left neighbour *razbrosat’ po vsemu* (‘scatter all over’ followed by a spatial location) having only 34 instances. Also, the sequence of words *po vsemu* is ambiguous, e.g. it can be a part of larger PPs, such as *po vsemu gorodu, domu, zalu* (over the whole city, house, hall), so from the viewpoint of automatic detection the MWE *sudja po vsemu* is more reliable.

Another example of an extended pattern is a complex reflexive expression: *drug druga* (‘each other’, lit. ‘friend friend-acc’), which is a multiword expression of its own, because no meaning of friendship is explicitly communicated here, as in *nenavidet’ drug druga* (‘to hate each other’, lit. ‘to hate friend friend-acc’). Even though the original pattern did not cover this structure, the expression has been detected for almost all prepositions in the form of PREP+*drug*-ending, because the reflexive expression allows the insertion of any preposition between the two elements, e.g. *drug k drugu* (‘to each other’, lit. friend to friend). Expressions of this sort resist the automatic identification by means of a simple pattern such as those used for other MWEs in the study.

It is well-known that ambiguity is abundant in natural languages. As discussed above, many word forms in Russian allow several morphological analyses and this applies to forms used in MWEs. Monolingual and bilingual dictionaries can also give an estimation of the semantic ambiguity by counting the number of senses and translations available for a word, though this will be the lower bound, because the number of senses and translations offered in dictionaries does not typically cover the full variety of types of possible uses: depending on a context, a word can be translated in many more ways than is suggested by a dictionary.

It was relatively straightforward to measure the reduction of morphological ambiguity. We can compare the number of morphological analyses before and after tagging of MWEs. The reduction of semantic ambiguity can be measured only indirectly by comparing the difference between the number of senses detected in a monolingual dictionary and the number of translations in a bilingual dictionary against the same numbers after tagging of MWEs, because we can assume that each MWE has only one sense, given the ‘one-sense-per-collocation’ hypothesis. Even in cases when the hypothesis does not hold, as in the case of the reflexive MWE *drug druga*, which can be translated in many different ways depending on the main predicate in a clause, the combination of the two words in an MWE saves from the possibility of their separate translation as *companion, friend, mate, pal, comrade, colleague, fellow*, etc.

Table 3 shows the level of the ambiguity in the original texts and the estimates for its reduction using the list of MWEs. The morphological analysis was performed using Mystem (Segalovich, 2003), a high-performance analyser which is also used in Yandex, a major Russian search engine. The results show that 41% of Russian word forms are ambiguous with respect to their morphological features with an average number of 4.6 analyses per ambiguous word (1.9 on average for all words).

The estimation of semantic ambiguity is based on electronic copies of the monolingual Ozhegov dictionary (Ozhegov, 1988) and the Oxford Russian bilingual dictionary (ORD, 2000). The former has 37785 entries with 1.6 senses per entry

	Morphology	Monolingual	Bilingual
Coverage	55022365	38508185	39056759
Average ambiguity	1.90	4.38	11.66
No of ambiguous words	22790728	19254090	19528375
Ambiguity per ambiguous word	4.59	8.76	23.32
Ambiguity after MWEs	4.06	8.39	21.72
Improvement	10.66%	4.27%	6.86%

Table 3: The analysis of the ambiguity resolution

on average, while the Russian-English part of the latter has 40303 entries with 1.9 translations per entry. The dictionaries were applied to simple tagging of the running text in the corpus, whereby every word listed in the dictionaries was tagged with the respective number of its senses and translations. The experiment also showed that either of the two dictionaries covers about 70% of the running text (noncovered words are typically proper names). Since more frequent words typically exhibit greater polysemy, the polysemy in the running text is larger. A word has about 4.4 senses on average according to (Ozhegov, 1988) and 11.7 translations according to (ORD, 2000). However, these counts are slightly misleading, because about half of the words in the corpus are not ambiguous. But if a word *is* ambiguous, it exhibits a much greater set of possible senses and translations: for instance, (ORD, 2000) lists the word *big* as having 35 translations in various contexts, so if the average ambiguity in the corpus is counted for ambiguous words only, it reaches 8.8 for senses and 23.3 for translations.

The results for morphological and semantic ambiguity are summarised in Table 3. After the application of the list of MWEs (they cover only about 2% of the total corpus size), the level of ambiguity for ambiguous lexical items goes down to 4.1 for morphological analysis, 8.4 for senses and 21.7 for translations. This gives a drop of about 11% for ambiguity in morphological analysis, 4% for ambiguity of senses and 7% for translations.

5 Conclusions

The paper reports the first attempt to apply computational methods to the detection and use of multiword expressions in Russian. The study resulted in a list of about 700 prepositional phrases which is available from `http://www.comp.leeds.ac.uk/ssharoff/frqlist/mwes-en.html`. The list offers rough results of MWE selection: it includes proper idioms, of the type one can find in a phraseological dictionary, in particular items missed or underdescribed in such dictionaries, so that it can be used as a source for improving them. However, it also includes items on the edge between idioms and other types of lexicalised phrases, for instance, grammatical constructions or institutionalised phrases.

The study shows that a simple method with little syntactic knowledge about the structure of PPs in Russian and no semantic resources can produce a useful list of MWEs. The combination of automatic detection of the most significant collocations and manual filtering of the results is not labour intensive and produces many expressions that are not covered in existing Russian dictionaries.

The next immediate step would be to use the lists for the study of translation equivalence between English and Russian, because MWEs are also not adequately represented in bilingual dictionaries, whereas their translation causes significant problems for language learners as well as for machine translation systems. For instance, the Oxford Russian Dictionary lists 13 translations of *bez* (without), including such idioms as *bez uma* ('be crazy about something', lit. 'without mind'), but fails to list many other more frequent constructions, such as *bez ocheredi* (to jostle to the front of the queue, lit. 'without queue'), *bez umolku* ([to talk] nonstop), *bez sleda* ([to vanish] without any hint), etc.

The lists can also act as a useful resource for morphological and semantic disambiguation. The list covers about 2% of the running text in the corpus, yet it reduces semantic ambiguity in the running text by 4–7%, and morphological ambiguity by 11%. We did not experiment with the reduction of syntactic ambiguity, because there is no Russian syntactic parser that can give robust parsing of an unrestricted corpus, such as that used in the study. Also, there is no easy way to force existing parsers to treat the identified MWEs as separate syntactic units on the clause level. However, we expect that accuracy will increase, because the set of identified MWEs reduces the number of PP attachment problems, as each MWE acts as an adjunct unit of its own within the clause.

The domain of prepositional phrases has been chosen specifically because it is relatively easy to guess the structure from the form by means of shallow parsing. Further experiments may consider detection of other types of MWEs, in particular, with light verbs, such as *brat' primer* (to follow the example of someone, lit. 'take example'), which are also very important for translation, but given the free word order in Russian this extension requires syntactic parsing to detect the dependency structure.

Acknowledgements

I'm grateful to Dmitrij Dobrovol'skij, Tony Hartley, Viktor Pekar, and especially to Bogdan Babych for useful discussions.

References

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–1940.

Dmitrij Dobrovol'skij. 2000. Contrastive idiom analysis: Russian and German idioms in theory and in the bilingual dictionary. *International Journal of Lexicography*, 13(3):169–186.

A.I. Fedorov, editor. 1995. *Frazeologicheskii slovar russkogo literaturnogo iazyka*. Nauka, Novosibirsk.

William Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proc. of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of COLING-ACL*, pages 483–490.

Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

Archibald Michiels and Nicolas Dufour. 1998. DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proc. of First International Language Resources and Evaluation Conference*, pages 1179–1186. Granada, Spain.

G. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4).

OED. 1989. *Oxford English Dictionary*. Clarendon Press, Oxford.

ORD. 2000. *The Oxford Russian Dictionary*. Oxford University Press, Oxford, 3rd edition.

S. I. Ozhegov. 1988. *Slovar' russkogo iazyka*. Russkii iazyk, Moskva, 20th edition.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, pages 49–56.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. Technical Report No. 2001-03, LinGO Working Paper.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proc. of MLMTA-2003*, Las Vegas.

Serge Sharoff. 2004. Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, and P. Rayson, editors, *Corpus Linguistics Around the World*. Rodopi, Amsterdam.

J. Zhang, J. Gao, and M. Zhou. 2000. Extraction of Chinese compound words – an experimental study on a very large corpus. In *Proc. of the 2nd Chinese Language Processing Workshop, ACL 2000*, pages 132–139.