Towards a Typology of Microsyntactic Constructions

Tania Avgustinova[1(✉)] and Leonid Iomdin[2(✉)]

[1] Department of Language Science and Technology, Saarland University, 66123 Saarbrücken, Germany
avgustinova@coli.uni-saarland.de
[2] Institute for Information Transmission Problems, Russian Academy of Sciences, 103051 Moscow, Russia
iomdin@iitp.ru

Abstract. This contribution outlines an international research effort for creating a typology of syntactic idioms on the borderline of the dictionary and the grammar. Recent studies focusing on the adequate description of such units, especially for modern Russian, have resulted in two types of linguistic resources: a microsyntactic dictionary of Russian, and a microsyntactically annotated corpus of Russian texts. Our goal now is to discover to what extent the findings can be generalized cross-linguistically in order to create analogous multilingual resources. The initial work consists in constructing a typology of relevant phenomena. The empirical base is provided by closely related languages which are mutually intelligible to various degrees. We start by creating an inventory for this typology for four representative Slavic languages: Russian (East Slavic), Bulgarian (South Slavic), Polish and Czech (West Slavic). Our preliminary results show that the aim is attainable and can be of relevance to theoretical, comparative and applied linguistics as well as in NLP tasks.

Keywords: Microsyntax · Typology · Comparative linguistics · Multilingual resources · Slavic languages · Russian · Bulgarian · Polish · Czech

1 Background

Written and spoken communication relies on a large amount of "prefabricated language"[1]. Many elements of this prefabricated language belong to what can be called syntactic idioms (Jackendoff 1997), the domain of microsyntax (Iomdin 2006, 2017; Apresjan et al. 2010). These microsyntactic elements can neither be handled within the lexicon alone nor interpreted compositionally by standard grammar rules. The syntactic behavior of phraseological units in a sentence, their lexical combinatorics, the communicative interaction with other elements of the discourse and even the used prosodic pattern can be very specific. No wonder: phraseology is the fragment of language in which ancient, long-established elements – syntactic constructions, lexical units, and grammatical forms come into close contact with modern language use, sometimes

[1] This term has been actively used since 1970s in language learning studies, e.g. (Hakuta 1974).

G. Corpas Pastor and R. Mitkov (Eds.): Europhras 2019, LNAI 11755, pp. 15–30, 2019.
https://doi.org/10.1007/978-3-030-30135-4_2

forming combinations so intricate that they can confuse not only a foreigner brilliantly speaking the language in question but also a well-educated native speaker. Moreover, it is not just the fact of a unit belonging to phraseology that makes it peculiar with regard to a free word combination but the fact that practically every such unit proves to be syntactically unique.

On the practical side of idiomaticity description (as seen from the viewpoint of text generation), we need to know what linguistic situations and phenomena may be conveyed by standard expressions, although alternatives are normally available, and to know what these alternatives are and which are preferable – regular or idiomatic. As stated by (Warren 2005) following (Sinclair 1991), language users have at their disposal a number of more or less pre-constructed phrases, so that the production of texts involves alternation between word-for-word combinations (open choice principle) and pre-constructed multi-word combinations (idiom principle). For the open choice principle, syntax is there to specify the slots into which memorized items – normally single words – can be inserted, while the idiom principle highlights the availability of memorized semi-pre-constructed combinations as single choices, even though they might appear to be analyzable into segments. What is more, (Mel'čuk 1996) suggests that memorized expressions outnumber single words. (Langacker 1998) makes a distinction between stored low-level patterns, many of which incorporate particular lexical items, and high-level schemas, which are general and productive patterns, but suggests that the low-level structures "do much, if not most of the work in speaking and understanding". Native speakers obviously know more than words and rules of how to put them together because of frequency effects: we naturally memorize what is repeated. Moreover, retrieving more or less ready-made combinations of words requires less mental effort than composing an utterance word for word (Wray 2002) – which, by the way, foreign language learners have to resort to.

The idiomaticity under investigation here goes beyond plain single-word cross-linguistic correspondences and takes various forms of semantic non-compositionality. The classical examples are quite heterogeneous – from regular form-meaning pairings (including proverbs, allusions and clichés) over so-called formal idioms (or partially lexicalized constructions with idiosyncratic meaning) to collocations (realizing combinatorial potentials of words). We accept a working definition proposed by (Čermák 2007): "The idiom is such a unique and fixed combination of at least two elements for which it holds that at least some of these do not function, in the same way, in any other combination or combinations of the kind, or occur in a highly restricted number of them, or in a single one only".

Thus, traditionally studied idioms can be functionally equivalent to major word classes like verbs (*rub someone´s nose in it, change horses in midstream*), nouns (*skeleton in the cupboard, an Indian summer*) or adverbs (*with hands down, in the middle of nowhere*). Other grammar idioms are equivalent to grammar words and used in the same function, e.g. English, Czech, or Russian prepositions (like *with regard to, in view of,* Ru. *в свете* 'in the light of', *за неимением* 'for want of', *Cz. na úkor* 'at the expense of', *s výjimkou* 'with the exception of'); conjunctions (Ru. *как будто* 'as though', *будь то ... или* '≈ be it… or'; Cz. *i když* 'even though'), particles (Ru. *только что* 'just now', *разве что* '≈ if any', Cz. *jen jestli* 'only if', *co kdyby* 'what

if'), pronouns (Ru. *что бы то ни было* 'whatsoever', Cz. *kdokoliv který* 'whoever') etc.

In the class of multiword prepositions, Čermák further distinguishes two types: (a) non-paradigmatic multiword prepositions, formed irregularly with the help of words belonging to various parts of speech, for which corpus-based lists can be created (e.g. Ru. *что касается* 'as concerns', *один на один с* 'one-on-one with', Cz. *co do* 'as for', *počínaje od* 'beginning with', *spolu s* 'together with', *vzhledem k* 'in view of', *tváří v tvář* 'face to face' – as in *tváří v tvář ženě* 'in the presence of the woman'), and (b) paradigmatic multiword prepositions, a potentially open class formed regularly with highly frequent nouns following a unified pattern [P N^{abstr} (P)] – e.g. Ru. [P N_{LOC}] *в интересах* 'in the interests of', [P N_{DAT} P] *по сравнению с* 'as compared to', [P N_{ACC} P] *в отличие от* 'in contrast to', Cz. [P N_{GEN}] *z hlediska* 'from viewpoint of', [P N_{ACC}] *pro případ* 'in case of', [P N_{LOC}] *v oblasti* 'in the field of'; [P N_{INSTR}] *pod vlivem* 'under the influence of'; [P_1 N P_2]: *na rozdíl od* 'unlike', *s ohledem na* 'with regard to', *ve srovnání s* 'in comparison with', etc.

As argued in (Sag et al. 2002), the enormous variety of multiword expressions (MWE) call for distinct treatment including (i) listing words with spaces, (ii) hierarchically organized lexicons, (iii) restricted combinatorial rules, (iv) individual lexical selection, (v) idiomatic constructions and (vi) simple statistical affinity. For the analyzed English data, the classification in Fig. 1 has been proposed.

Lexicalized MWE have at least partially idiosyncratic syntax or semantics and, with regard to "decreasing lexical rigidity" can be further broken down into fixed (i.e. fully lexicalized and undergoing neither morphosyntactic variation nor internal modification), semi-fixed (undergoing some degree of lexical variation) and syntactically-flexible expressions (exhibiting a much wider range of syntactic variability). Institutionalized MWE are syntactically and semantically compositional but "statistically idiosyncratic", occurring with "markedly high frequency" in certain contexts. Importantly, the conclusion drawn in Sag's work is relevant for the purpose of this paper: "Scaling grammars up to deal with MWEs will necessitate finding the right balance among the various analytic techniques. Of special importance will be finding the right balance between symbolic and statistical techniques".

The creation of multilingual microsyntactic resources can be useful in a variety of research and development projects, from computer-assisted language learning tools to cross-linguistic studies including typological and cross-cultural dimensions. One important project that could benefit from such resources is Universal Dependencies community (http://universaldependencies.org) especially if – as required by (Croft et al. 2017) – typological research on language universals is systematically considered and dependencies are based on universal construction types over language-specific strategies. Hence, a universal annotation scheme would be able to use a classification of constructions as its universal foundational layer, avoiding solutions reliant on language-specific strategies while capturing the most commonly occurring strategies, too.

On a deeper systematic level, the formalism of lexical functions and paraphrasing rules (Melčuk 1998; Wanner 1996) is worth considering for certain classes of phenomena, too. The lexical functions cover both semantic-derivation relations (synonyms, antonyms, conversives, nominalizations, verbalizations, actant names,

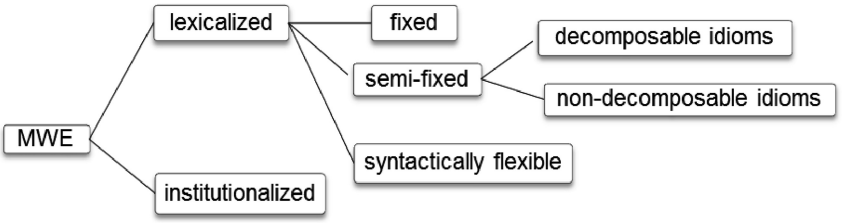

Fig. 1. Classification of multiword entities

adjectives characterizing actants, etc.) and collocational relations (intensifiers, positive and negative evaluators, light verbs, realization verbs, etc.). Paraphrasing rules, as understood by the Meaning-Text Theory, are formulated in terms of lexical functions and are applicable both within a language and between languages, i.e. as intra-lingual and inter-lingual paraphrasing. In particular, paraphrasing rules may be very convenient for rule-based machine translation systems.

There have been relatively few studies specifically devoted to the typology of MWE units so far, even though certain aspects of this typology (or, at least, comparison of two or more languages) were considered in a number of research studies, starting from (Blanco 1997), who discusses the typology of translation divergence in compound nouns between French and Spanish), and in a computer-assisted language learning project CALLLex, where the formalism of lexical functions was used to compare multiword entities involving lexical functions in Russian, German, French, and, later, Spanish (Apresjan et al. 2002; Boguslavsky et al. 2006).

The project considered here, however, is the first attempt at creating a typology of microsyntax (i.e. nonstandard syntactic constructions and syntactic idioms) for a number of Slavic languages.

2 Microsyntax Cross-Linguistically

Multi-component units are characterized by language-specific idiomaticity involving various degrees of non-compositionality due to the grammaticalization and/or lexicalization of the respective expression. The easily observable fact that cross-linguistic equivalents of such units belonging to a particular language usually appear as multi-component units in other languages, too, suggests that microsyntactic phenomena are cross-linguistically comparable, especially when closely related languages are considered.

Microsyntactic units defy uniform interpretation while exhibiting little freedom in formation and pronounced anomaly in structure. Yet, as far as cross-linguistic comparison is concerned, many of the peculiarities of microsyntactic units of one language are, at least partially, reproducible in cognate languages. This fact makes it a feasibly task to build a typology of microsyntactic phenomena starting from a particular language, for which the classification has been (more or less) established, and finding the equivalents of these phenomena in other languages and, possibly, subgroups of languages. Specifically, we start with Russian, for which microsyntactic research has been

going on for almost two decades and the respective microsyntactic dictionary and corpus resources have been developed, and build our typology for several Slavic languages, taking Russian as the pivot language. Altogether, we focus on four Slavic languages representing the three main sub-groups of the language family – Russian (East Slavic), Bulgarian (South Slavic), Polish and Czech (both West Slavic). This language selection provides us with the required typological variability in grammar. In addition, all four languages are well-resourced: large-scale national corpora are available as well as parallel multilingual data and dedicated query tools. For an adequate comprehensive analysis of microsyntactic phenomena from a typological perspective we employ both symbolic and statistical techniques.

By methodically studying and elaborating the inventory of constructions established on the basis of Russian, we aim at a fine-grained and cross-linguistically applicable hierarchy of microsyntactic phenomena. Lexical, grammatical and constructional information about the monolingual microsyntactic units has to be captured in a modular way in order to enable their language-family oriented interpretation. Cross-linguistic correspondences of microsyntactic phenomena need to be found, investigated and accordingly classified. These include bilingual correspondences between Russian and another Slavic language, as well as multilingual correspondences. Our objective is to design a comprehensive typology resource encompassing the multitude of conventionalized multi-word combinations cross-linguistically in the form of a core Slavic micro-syntactic database to be used for educational purposes and in language-technology applications.

The central linguistic resource we use is the Russian National Corpus (RNC), in particular, its main sub-corpus (over 600,000,000 tokens), the parallel corpus with counterparts of Russian texts in five Slavic languages, i.e. Belarussian (9,500,000 tokens), Bulgarian (3,800,000 tokens), Czech (1,500,000 tokens), Polish (6,300,000 tokens) and Ukrainian (9,300,000 tokens), as well as the syntactically annotated sub-corpus (1,200,000 tokens). Two different interfaces of the main RNC sub-corpus and the parallel corpora are freely available through any web browser at http://ruscorpora.ru and http://corpus.leeds.ac.uk/ruscorpora.html. The former interface provides access to the whole array of the main RNC sub-corpus, while the latter provides access to a 50,000,000-token fragment of this sub-corpus and some other sub-corpora but it also enables searching other Russian corpora, the internet, or performing a combined search. Additionally, in the newest version of the syntactically annotated aub-corpus SynTagRus, over 10,000 sentences have microsyntactic tags annotations, and the inventory of microsyntactic elements present in SynTagRus counts about 1,100 items.

We started with the collections of multiword expressions provided at the RNC website (http://www.ruscorpora.ru/obgrams.html), which were selected from RNC frequency collocation database and supplemented with the data from the Malyj Akademičeskij Slovar' (MAS 1999) and a collocation dictionary (Rogozhnikova 2003):

(i) Prepositions (http://ruscorpora.ru/obgrams-PR.html)
(ii) Adverbial and Predicatives (http://ruscorpora.ru/obgrams-ADV.html)
(iii) Parenthetical expressions (http://ruscorpora.ru/obgrams-PARENTH.html)
(iv) Conjunctions (http://ruscorpora.ru/obgrams-CONJ.html)
(v) Particles (http://ruscorpora.ru/obgrams-PART.html)

These inventories are used for finding translational equivalents in parallel and multilingual sub-corpora.

3 Empirical Foundation

Based on the Russian data, three major types of microsyntactically relevant material can be preliminarily distinguished and will be referred here – for the sake of illustration – as lexically idiomatic Type A, syntactically idiomatic Type B and constructionally idiomatic Type C.

3.1 Lexically Idiomatic Cross-Slavic Correspondences (Type A)

Type A contains lexicalized multi-component units with an idiosyncratic meaning and word-like status, e.g., compound prepositions (*в отличие от* 'unlike', *на тему* 'on the subject of', *по причине* 'due to'), or closed-class combinations of pronominal and function words (*как будто* 'as though', *будто бы* 'as if', *разве что* 'perhaps only', *только что* 'just now', *что за* 'what the', *не прочь* 'don't mind'), as well as parenthetical expressions (*стало быть* 'so then', *была не была* 'nothing to lose', *между прочим* 'by the way') or semi-lexicalized patterns (*кто/.../что угодно* 'whatever/whoever', *чёрт/.../бог знает кто/.../зачем* 'Devil/…/God knows why'). Three subtypes of Type A constructions are distinguished:

Type A1: Multiword Idiomatic Prepositions
Typically the pattern [P_1 N^{abstr} (P_2)] is followed, as summarized in Table 1. Multilingual correspondences are illustrated in Table 2.

Table 1. Russian idiomatic multiword prepositions

P_1	$N^{abstract}$	(P_2)	N_{case}	English equivalent	Available alternatives	Examples of alternative implementations
в	*связи*	*с*	N_{instr}	with regard to	*в DEM связи*	*в этой связи* 'in this regard'
в	*отличие*	*от*	N_{gen}	unlike, in contrast to		
в	*адрес*		N_{gen}	in/to the address of	*в INT/POSS адрес*	*в мой адрес* 'to my address, toward me'; *в чей адрес* 'to whose address'
во	*время*		N_{gen}	at the time of	*в DEM время*	*в то же время* 'at the same time'
по	*причине*		N_{gen}	because of, due to	*по DEM/INT причине*	*по какой причине* 'for what reason'
в	*соответствии*	с	N_{inst}	in accordance with		
на	*тему*		N_{gen}	on the topic of	*на DEM/INT тему*	*на какую тему* 'on which subject'

For some of the compound prepositions, alternative realizations exist, involving demonstrative (DEM), interrogative (INT) or possessive (POSS) modifiers of the nominal component (cf. Type A1). There are obvious parallels to what (Čermák 2007) considers as grammar idioms which are equivalent to grammar or auxiliary words and therefore used in the same function.

Table 2. Cross-lingual correspondences to Russian idiomatic multiword prepositions

Russian	Bulgarian	Polish	Czech
в связи с	*във връзка с*	*w związku z*	*v souvislosti s*
в отличие от	*за разлика от*	*w przeciwieństwie do*	*na rozdíl od*
в адрес	*на адреса на*	*na adres*	*na adresu*
во время	*по време на*	*w czasie*	*během*
по причине	*поради*	*z powodu*	*z důvodu*
в соответствии с	*в съответствие със*	*zgodnie z*	*v souladu s*
на тему	*на тема*	*na temat*	*na téma*

Alternative implementations, e.g. Polish constructions allowing for modifiers like *w tym związku* 'in this respect', *na pański adres* 'to your address', *z jakiego powodu* 'for which reason', *na ten temat* 'on this topic', have to be included into the database with details of their syntactic and semantic behavior too.

Type A2: Combinations of Closed Class Pro-forms and Function Words
The examples are given in Table 3. Again, the intended multilingual output is illustrated in Table 4. For brevity, only several examples are given.

Table 3. Russian syntactic idioms acting as function words

adv/prep/prt	aux_1	pron/conj	prt	$adv/A/N^{abstr}$	aux_2	English equivalents
		как	*бы 1*			'as if, like, sort of' (particle): *Он как бы играл* 'he sort of played'
		как	*бы 2*			'lest' (conjunction):*Я боюсь, как бы он не опоздал* 'I fear lest he should be late'
		всё	*же*			'all the same'
только		*что*				'just now'
		тем	*не*	*менее*		'however', 'yet'
между		*тем*				'meanwhile'
между				*прочим*		'by the way'
		как	*буд-то*			'as though'
		что	*ли*			'or something'
		что	*за*			'what kindof a'

(*continued*)

Table 3. (*continued*)

adv/prep/prt	aux_1	pron/conj	prt	adv/A/N^{abstr}	aux_2	English equivalents
		тот	*же*			'same as'
пока		*что*				'as yet'
разве		*что*				'perhaps only'
		тем		*более*		'especially'
			не	*прочь*		'don't mind'
		то	*и*	*дело*		'every now and then'
		так	*и*		*быть*	'so be it'
		всё		*равно*		'just the same'
		как		*раз*		'just, exactly'
не		*тут-*	*-то*		*было*	'nothing of the kind'
	будь	*что*			*будет*	'whatever happens'
	была		*не*		*была*	'I'll risk it; nothing to lose'
	стало				*быть*	'hence'

Table 4. Cross-lingual correspondences to Russian syntactic idioms acting as function words

Russian	Bulgarian	Polish	Czech
как бы 1	*като че ли*	*jakby*	*jakoby*
всё же	*все пак*	*jednak*	*přesto, stejně*
только что	*току-що*	*dopiero*	*zrovna, právě*
тем не менее	*въпреки това*	*tym niemniej, mimo wszystko*	*nicméně, presto*
между тем	*същевременно*	*tymczasem*	*mezitím*
между прочим	*между другото*	*nawiasem mówiąc, przy okazji*	*mimochodem*

The result of a sample search for the grammar (conjunction-like) idiom *как будто* ('as though') in the four selected languages is provided below to exemplify the procedure. It includes translations of a sentence from Mikhail Bulgakov's *Master and Margarita* into the three Slavic languages. The English equivalent of the sentence is *He looked Berlioz up and down* ***as though*** *he were measuring him for a suit*.

Ru: *Он смерил Берлиоза взглядом,* ***как будто*** *собирался сшить ему костюм*
Bg: *Той измери Берлиоз с поглед,* ***като че ли*** *му взимаше мярка за костюм*
Pl: *On zmierzył Berlioza spojrzeniem,* ***jakby*** *zamierzał uszyć mu garnitur*
Cz: *Změřil si ho pátravým pohledem,* ***jako by*** *se mu chystal šít oblek*

Type A3: Semi-lexicalized Patterns with Constant and Variable Parts
These are syntactic idioms in which some parts are fixed whilst others may vary: the extent of variation can be different and sometimes extremely difficult to generalize. So, in the two examples given in Table 5, one part is represented by an interrogative pronoun while another is instantiated by a concrete word (*угодно* ≈ 'ever' in (a) and *знает* 'knows' in (b), to account, respectively, for expressions like *Поеду куда угодно*

'I'll go anywhere' and *Чёрт знает, что он замышляет* 'The devil only knows what he is up to'.

Table 5. Russian semi-lexicalized patterns

	$N^{variable}$	$V^{constant}$	Interrogative pro-form	$Adv^{constant}$	English equivalent
(a)			*кто/что/где/зачем/куда*	*угодно*	'wh-ever'
(b)	*чёрт/бес/бог*	*знает*	*кто/что/где/зачем/куда*		'God knows'

The syntactic idiom (b) allows for a lexical variation with the noun being *чёрт* 'devil', *бес* 'demon', but also *бог* 'God', *Аллах* 'Allah', as well as several other names for devils and deities (but not all!) and bizarrely enough, *нёс* 'dog'. Note that cross-linguistic typological research of such patterns is quite demanding with regard to time, effort, and qualification.

3.2 Syntactically Idiomatic Cross-Slavic Correspondences (Type B)

A separate class consists of syntactic idioms that have neither structural transparency nor word-like status. One of their key features is that they acquire valence properties as a unit, cf. (a) *как быть* (as in *Как быть профессору* [X] *со студентами* [Y] *на экзамене* [Z], *если они списывают?*) 'What should the professor do (about the students if they cheat in the exam?), and (b) *то ли дело,* constructing a contrapositive "*не* V_{fin} …, *то ли дело* X" (*Он не любит делать уроки, то ли дело мультики смотреть* 'He does not like doing homework – how much better is watching cartoons'; *С детьми ты не играешь, то ли дело одноклассники в интернете*. 'You don't play with the kids, but with classmates on the internet it's a different story'). This class further includes (c) the transitive use of the finite forms of the verb БЫТЬ 'be' (only in the indicative future: *буду, будешь,* etc.), selecting a complement with food/drink/smoking semantics in a situation of immediacy and service (*буду только чай* 'I'll have just tea', *торт не буду* 'I don't want cake', *Он будет прямо из бутылки* 'he'll drink straight from the bottle', *Будешь сигарету?* 'Will you have a cigarette'). Valence structures of Type B syntactic idioms are summarized in Table 6.

The number of weakly lexicalized syntactic idioms of this type is much smaller than that of lexicalized ones of Type A1, and we expect them to be more language specific and not always allow for close equivalents even in related languages. The outcome of a sample search for the syntactic idiom (a) in available Slavic languages is illustrated in Table 7 to give us an idea of the nature of the variation we are confronted with. Only one of the three Russian translations of Lewis Carroll's *Alice in Wonderland* (Russian_2, by Vladimir Nabokov) uses this syntactic idiom, and there is no equivalent of the verb *быть* in the other Slavic translations, which is not surprising since all translations were made from the English original, rather than Russian. A quick search through Russian-Polish parallel corpus of RNC for the equivalents of *как быть* reveals a similar result. There is only one occurrence of the Polish verb *być* in the pair Pl. *A co będzie z tymi, co już byli?* – Ru. *А как быть с теми, кто был раньше?,* which

Table 6. Russian syntactic idioms

(a)

как быть	**X_{dat}**	**$Y_{[with]}$**	**$Z_{[situation]}$**
'what to do about'	профессору	со студентами	на экзамене
	'professor'	'with students'	'at examination'

(b)

не V_{fin} …,	*то ли дело*	**X**
(negated proposition)		(non-negated proposition)

(c)

[situation: immediacy + service] BE-auxiliary[future] [transitive]	**$X_{[food/drink/tobacco]}$**
буду/будешь/будет/будем/будете/будут	*чай*
'will have'	'tea'

however has a notable difference in meaning: the Polish sentence simply asks what will happen to those who were before, while the Russian translation asks what the protagonist should do about these people. The Russian/Bulgarian and Russian/Czech parallel corpora provided no such results, either. Yet the two East Slavic languages showed a very close correspondence for this syntactic idiom: *як быць* in Belarussian and *як бути* in Ukrainian in almost all cases.

Table 7. Cross-lingual correspondences to Russian syntactic idioms

English	Alice had no idea what to do,
Russian	Алиса растерялась.
Russian_2	Аня не знала, как ей быть.
Russian_3	Бедная Алиса не знала, что ей делать;
Ukrainian	Аліса не знала, що робити;
Belarussian	Алеся ня мела ніякага ўяўленьня, што рабіць.
Polish	Alicja nie wiedziała, jak wybrnąć z tej sytuacji.
Polish_2	Zupełnie nie wiedząc, co począć,
Czech	Alenka nevěděla, co počít,
Slovak	Alica raz nevedela, čo robiť;
Slovene	Alica sploh ni vedela, kaj naj stori;
Croatian	Jadna Alica! Šta će sad? Ne znajući kako da se izvuče iz neprilike
Serbian	Алиса није имала појма шта да ради.
Macedonian	Алиса воопшто не знаеше што да прави,
Bulgarian	Алиса не знаеше какво да стори

3.3 Moderately Transparent Cross-Slavic Correspondences (Type C)

The number and the variety of moderately transparent non-standard syntactic constructions are quite impressive (Fig. 2). In many cases they have no counterparts even in closely related languages.

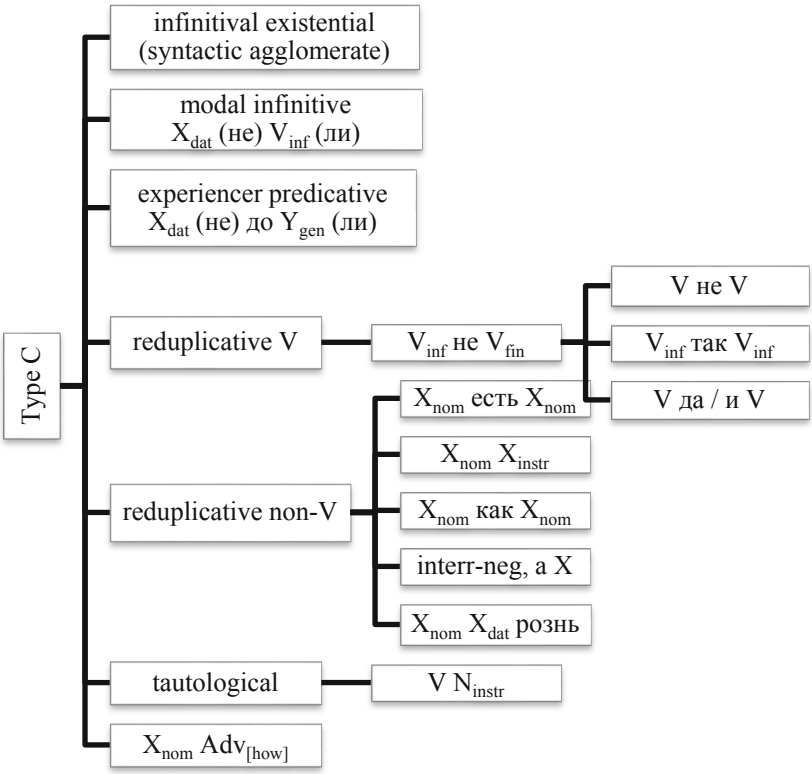

Fig. 2. Russian moderately transparent non-standard syntactic constructions

Relevant examples include the modal infinitive with a dative subject (мне скоро улетать 'I'm leaving [lit. flying away] soon'), reduplicative/repetitive expressions (знать не знаю, но… 'as for knowing, I don't know, but…'; люди как люди 'just the usual people [lit. people like people]'; гулять так гулять 'let's celebrate properly [lit. celebrate so celebrate]'), expressions with constant and variable parts (какой-никакой, а X – cf. Здесь я какой-никакой, а герой 'Here I'm a hero in any case'; or X-у не до Y-а – as in Им не до сна 'They have more important things than sleeping, they don't feel like sleeping').

The syntactic agglomerate construction (Fig. 3) has been first studied by (Apresjan and Iomdin 1989) and later formalized in (Avgustinova 2003). It poses a theoretical

challenge to lexicon and grammar. The respective equivalents from a cross-linguistic perspective are quite diverse in lexical realization and with regard to structure.

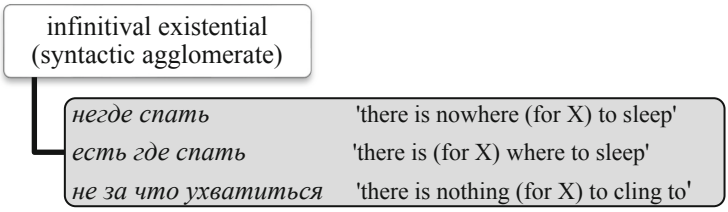

Fig. 3. Russian syntactic agglomerate constructions

The next two constructions (Fig. 4) display the dative realization of the subject.

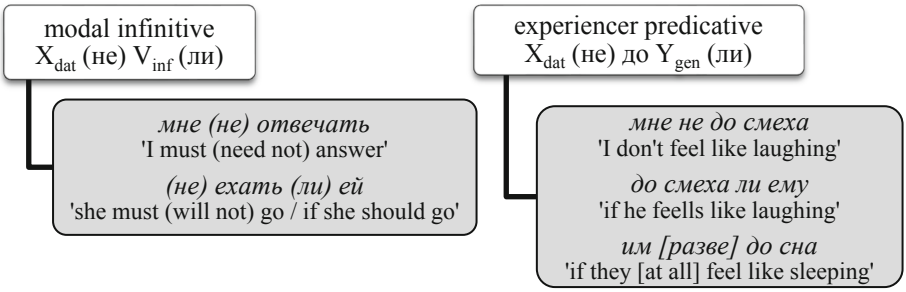

Fig. 4. Dative realization of the subject valence

Reduplicative or repetitive expressions (Fig. 5) can be verbal and non-verbal.

The reduplication in tautological constructions is of a different nature, for which in Russian the pattern (Fig. 6) is found.

Finally, colloquial expressions relating to personal circumstances follow the pattern in (Fig. 7).

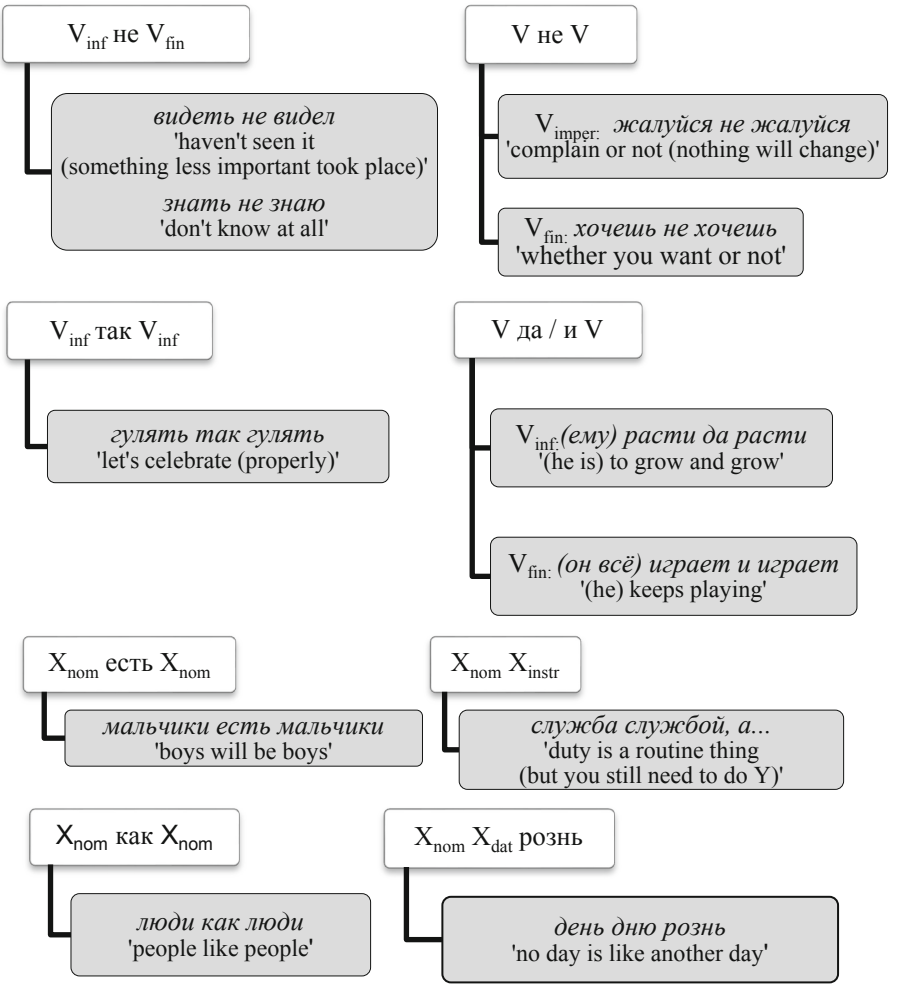

Fig. 5. Reduplicative expressions

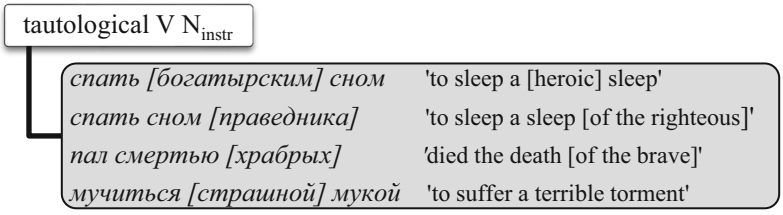

Fig. 6. Tautological constructions

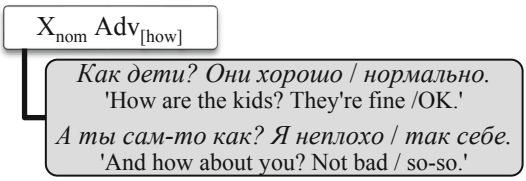

Fig. 7. Colloquial constructions

Let us consider the last type of reduplicative non-V constructions in Fig. 5: it manifests itself in expressions like *X X-у рознь* ≈ 'one X is different from another X', where the noun *рознь* has a separate meaning ≈ 'difference' which does not appear in any other context. Strange as it may seem, this extraordinary construction can be found in other Slavic languages, somewhat in a different form but retaining the repetition. The examples in Table 8 are taken from the parallel sub-corpora of the RNC.

Table 8. Cross-lingual correspondences

Russian: Bulgarian:	*Агент агенту рознь.* *Има агенти и агенти.* 'There are different sorts of agents'
Russian: Belorussian:	*Грешки грешкам — рознь* (Гоголь) *Грашкі на грашкі ня выходзяць.* 'Little sins can be different'
Russian: Ukrainian:	*Погода погоде рознь, да и день — дню!* *Верем'є верем'ю нерівне, та й днина днині!* (Мартович) 'Weather is different from weather, and one day is different from another day'
Russian: Polish:	*Оказывается, пуля пуле рознь.* *Kula, jak się okazywało, kuli nie była bynajmniej równą* (Sapkowski) 'It turns out one bullet is different from another'

4 Conclusion

We have presented the initial stage of an international collaboration for creating a typology of microsyntactic phenomena on the basis of contrastive syntactic and lexicographic studies. The empirical base is constituted by Slavic languages as they are mutually intelligible to various degrees. In such a setup, monolingual idiomaticity of microsyntactic constructions can be approximated cross-linguistically from the perspective of its comprehensibility to speakers of closely related languages. The presented approach can be instrumental in the creation of monolingual, bilingual and multilingual resources that deal with non-standard syntactic phenomena and thus promising in improving natural language processing applications.

Acknowledgements. The authors are grateful to the Russian National Foundation (grant No. 16-18-10422-P) and the German Science Foundation (DFG, grant within the Collaborative Research Centre SFB 1102) for their partial support of this research.

References

Apresjan, J.D., Boguslavsky, I.M., Iomdin, L.L., Sannikov, V.Z.: Theoretical problems of russian syntax. Interaction of the grammar and the lexicon. [Teoretičeskie problemy russkogo sintaksisa]. In: Apresjan, J.D. (ed). Jazyki slavjanskix kultur Publishers, Moscow, 408 p. (2010). ISBN 978-5-9551-0386-0. (in Russian)

Apresjan, J.D., Boguslavsky, I.M., Iomdin, L.L., Tsinman, L.L.: Lexical functions in NLP: possible uses. Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday, Peter Lang, pp. 55–72 (2002)

Apresjan, J.D., Iomdin, L.L.: The construction of the NEGDE SPAT' type: syntax, semantics, lexicography. [Konstrukcija tipa NEGDE SPAT': sintaksis, semantika, leksikografija]. Semiotika i informatika, pp. 34–92. Vsesojuznyj institut nauchnoj i texnicheskoj informacii, AN SSSR, Moscow (1989). (in Russian)

Avgustinova, T.: Russian infinitival existential constructions from an HPSG perspective. In: Kosta, P. et al. (eds.) Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages, pp. 461–482. Peter Lang Europäischer Verlag der Wissenschaft (2003)

Boguslavsky, I., Dyachenko, P., Barrios Rodríguez, M.A.: CALLEX-ESP: a software system for learning Spanish lexicon and collocations. In: Current Developments in Technology-Assisted Education. Badajos (Spain): FORMATEX, vol. 1, pp. 22–26 (2006)

Čermák, F.: Grammatical Idioms. Philologica Pragensia, XVII, vol. 2, pp. 75–90 (2007)

Croft, W., Nordquist, D., Looney, K., Regan, M.: Linguistic typology meets universal dependencies. In: Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), pp. 63–75 (2017)

Hakuta, K.: Prefabricated patterns and the emergence of second language acquisition. Lang. Learn. **24**(2), 287–297 (1974)

Iomdin, L.L.: Polysemous syntactic idioms: between the vocabulary and the syntax [Mnogoznačnye sintaksičeskie frazemy: meždu leksikoj i sintaksisom]. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialog-2006", Moscow, RGGU Publishers, pp. 202–206 (2006). (in Russian)

Iomdin, L.: Between the syntactic idiom and syntactic construction. Nontrivial cases of microsyntactic ambiguity. [Meždu sintaksičeskoj frazemoj i sintaksičeskoj konstruktsiej. Netrivial'nye slučai mikrosintaksičeskoj neodnoznačnosti]. In: SLAVIA, časopis pro slovanskou filologii, ročník 68, sešit 2–3, pp. 230–243 (2017). (in Russian)

Jackendoff, Ray: Twisting the night away. Language **73**, 534–559 (1997)

Langacker, R.W.: Indeterminacy in semantics and grammar. Paper presented to the Estudios de Lingüística cognitiva, 1998 (1998)

MAS: The Small Academic Dictionary of Russian in 4 volumes, A. P. Evgenyeva, ed. [Slovar' russkogo jazyka v 4-x tomax, MAS, Malyj Akademičeskij Slovar. Russkij Jazyk Publisher, Moscow (1999). http://feb-web.ru/feb/mas/mas-abc/default.asp

Mel'čuk, I.: Lexical functions: a tool for the description of lexical relations in a lexicon. In: Lexical Functions in Lexicography and Natural Language Processing, vol. 31, pp. 37–102 (1996)

Mel'čuk, I.: Collocations and lexical functions. In: Phraseology. Theory, Analysis, and Applications, pp. 23–53 (1998)
Rogozhnikova, R.P.: An Explanatory Dictionary of Collocations Equivalent to Words [Tolkovyj slovar sočetanij, ekvivalentnyx slovu]. Moscow, Astrel, 414 p. (2003). (in Russian.)
Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: a pain in the neck for NLP. Paper presented to the International Conference on Intelligent Text Processing and Computational Linguistics (2002)
Sinclair, J.: Corpus, concordance, collocation. Oxford University Press (1991)
Wanner, L.: Lexical functions in lexicography and natural language processing. John Benjamins Publishing (1996)
Warren, B.: A model of idiomaticity. Nordic J. Engl. Stud. **4**, 35–54 (2005)
Wray, A.: Formulaic language in computer-supported communication: theory meets reality. Lang. Awareness **11**, 114–131 (2002)