Yevgen Matusevych, Ad Backus, Martin Reynaert. Do we teach the real language? An analysis of patterns in textbooks of Russian as a Foreign Language. Published in *Dutch Journal of Applied Linguistics, 2*(2), 224--241. DOI: http://dx.doi.org/10.1075/dujal.2.2.07mat

Draft version: August 2012.

This article is about the type of language that is offered to learners in textbooks, using the example of Russian. Many modern textbooks of Russian as a Foreign Language aim at efficient development of oral communication skills. However, some expressions used in the textbooks are not typical for everyday language. We claim that textbooks' content should be reassessed based on actual language use, following theoretical and methodological models of Cognitive and Corpus Linguistics. We extracted language patterns from three textbooks, and compared them with alternative patterns that carry similar meaning by (1) calculating the frequency of occurrence of each pattern in a corpus of spoken language, and (2) using Russian native speakers' intuitions about what is more common. The results demonstrated that for 35 to 53 percent of all the recurrent patterns in the textbooks better alternatives could be found. We further investigated the typical shortcomings of the extracted patterns.

Keywords: second language teaching, communicative language teaching, Russian as a foreign language, cognitive linguistics, construction grammar, corpus linguistics.

1. Introduction

In the field of teaching Russian as a Foreign Language (RFL), the number of available textbooks has considerably increased during the last decade. Many of them claim to be based on the method of communicative language teaching (CLT), following the international trend. CLT emerged in the 1960s–1970s as an alternative for the traditional methods influenced by structural linguistics—Audiolingual Method and Situational Language Teaching. While structural methods focus on grammatical competence and attend to structure and form, CLT emphasizes the importance of communicative competence and focuses on communicative efficiency rather than on correctness in form (Richards & Rogers, 2001; Savignon, 2002; Spada, 2007). Foreign language, it is argued, is acquired *for* communication and *through* communication. One of the principles that CLT proclaims is that learners have to be taught appropriate language (e.g., Richards & Rogers, 2001; Johnson & Johnson, 1998), meaning that the language offered has to match what Hutchinson and Waters (1987) call *target needs*—that is, the situation in which learners will use the language.

In our case we have to consider the goals of CLT in the specific learning situations in which RFL textbooks are normally used. CLT aims at the development of oral communication skills, so the learning goal for students is to be able to communicate efficiently. Furthermore, students (at least at the breakthrough level) predominantly use Russian in everyday situations when engaging in informal communication with Russian native speakers (NSs). Now, consider the following example of a short dialog from a breakthrough-level RFL textbook (Korchagina & Stepanova, 2010):

(1a) *– Давай познакомимся: меня зовут Виктор.*
davay poznakomimsya menya zovut Viktor.
let.2S make.acquaintance:1PL(FUT) I.ACC call:3PL Viktor
'Let's get acquainted. My name is Viktor.'

– Очень приятно. А меня зовут Елена.
ochen' priyatno a menya zovut Yelena.
very pleasant and I.ACC call:3PL Yelena
'Nice to meet you. And my name is Yelena.'

This dialog is intended to introduce an *informal* way to get acquainted. Although all the expressions in (1a) are grammatically correct, some of them are not typical for this context. *Menya zovut* is a rather redundant phrase in colloquial language, and it certainly will not be used in two consecutive turns; and *davay poznakomimsya* tends to be used in its (contracted) question form: *poznakomimsya?*. A more typical rendition of what NSs would say in this informal situation could be (1b):

(1b) *– Познакомимся? Я Виктор.*
poznakomimsya ya Viktor
make.acquaintance:1PL(FUT) I Viktor
'Let's get acquainted? I'm Viktor.'

– Очень приятно. Елена.
ochen' priyatno Yelena
very pleasant Yelena
'Nice to meet you. Yelena.'

The expressions in (1b) are colloquial, concise and syntactically incomplete. However, the dialog is not more difficult to understand, it contains fewer lexical items (and thus is easier to learn), and, most importantly, it fulfills NSs' expectations.

Textbook authors have to make many such choices, of whether or not to include one or another expression. Unfortunately, these choices are not always dictated by actual language, while the goals of CLT and the theoretical assumptions of Cognitive Linguistics, the branch of linguistics closest to CLT (see Broccias, 2008), advocate they should be. We will argue that communicative breakthrough-level RFL textbooks tend to overuse formal language, which conflicts with the learners' goal to communicate in basic everyday situations. Thus, we believe that the field of RFL, and by extension language teaching in general, will benefit if we analyze at which points

communicative breakthrough-level RFL textbooks do not use natural colloquial language. Such evaluation is best based on authentic spoken language that can be found in respective corpora (see Gilmore, 2007; Barbieri & Eckhardt, 2007). Cognitive Linguistics provides a useful methodological framework for such analysis (see, e.g., Meunier, 2008).

In the present study we try to estimate to what extent patterns used in breakthrough-level RFL textbooks correspond to those in actual language use. Section 2 provides an overview of related work as well some important theoretical notions of Cognitive and Corpus Linguistics that have motivated our study. In section 3 we present our methodology and results, the implications of which are discussed in the concluding section 4.

2. CLT, corpora and Cognitive Linguistics

2.1 'Real' language in CLT

The idea of applying knowledge about 'real' language to CLT is in no way new. We see two main trends.

Re-inventing grammars with corpora. Carter and McCarthy (1995) claim that many communicatively-oriented ESL textbooks are based on the grammar of written English, while conversational patterns would be more beneficial. Furthermore, based on their corpus study they conclude that "even very small corpora … can yield recurring patterns of grammar that are not fully ... described in conventional grammars" (p. 154). Other corpus studies also address this problem (e.g., Glisan & Drescher, 1993; Biber & Reppen, 2002; see an overview by Barbieri & Eckhardt, 2007), and the approach has motivated new grammar models (starting from Hornby, 1954) and textbooks (e.g., Carter, Hughes, & McCarthy, 2000; McCarthy, 2004).

One of the most ambitious works in this area is *Collins COBUILD Grammar Patterns* (see Francis, Hunston, & Manning, 1996). Hunston and Francis (2000) describe the theory and practice behind it, and show how common patterns were detected using concordance searches. The resulting patterns in their grammar book are very abstract and mostly consist of word-class labels and function words (e.g., '**V**' for a verb, '**v**' for a verb group, etc.): '**V n over n**', '**it V to N**', '**MODAL inf than/as inf**', etc.

We find this methodology useful in terms of the very notion of pattern. However, for our purposes their definition is too abstract: it is unlikely that such patterns can be more helpful to the learner than a traditional grammar rule. Langacker (1987) and many other since argue that people store patterns that contain labels as well as patterns with specific (lexical) elements. We will elaborate on this, and provide our definition of 'pattern' in section 2.2, but it is important to stress that schematic (e.g., **N**) and specific (e.g., '**weather**') elements can compete in two related patterns

(i.e.: both '**The weather is ADJ**' and '**The N is ADJ**' are patterns). This understanding is further informed by the second trend.

Interest in formulaic language. Applied linguistics has witnessed a shift from an emphasis on grammar towards an emphasis on lexis in language teaching. The COBUILD approach reflects this trend too. Hunston and Francis (2000) note that corpora observations "blur the traditional distinction between lexis and grammar" (p. 250), mostly because the underlying dimension of frequency is important for both. Willis (1990) describes this methodology as "the lexical approach", and it is the basis of another work originating from COBUILD—*Collins COBUILD English Course*.

The lexical approach claims that "focus on communication necessarily implies increased emphasis on lexis, and decreased emphasis on structure" (Lewis, 1993, p. 33). An implication of this is that so-called 'formulaic language' becomes important. Ellis and Cadierno (2009) say that "there has never been more interest in second language phraseology" (p. 114), and list various terms that address similar phenomena: holophrases, prefabricated routines and patterns, formulaic speech, memorized sentences and lexicalized stems, lexical phrases, formulas, chunks, multi-word expressions and constructions. All these blur the boundary between syntax and lexicon, a stance clearly related to the development of Cognitive Linguistics (see Section 2.2).

CLT theorists claim that "the primary function of language is to allow interaction and communication", that "the structure of language reflects its functional and communicative uses" (Richards & Rogers, 2001, p. 161), that "language is seen as a social tool which speakers use to make meaning", and that "diversity is recognized and accepted as part of language development and use" (Berns, 1990, p. 104). These underpinnings make the theoretical base of CLT compatible with the usage-based model that underlies much of Cognitive Linguistics. Broccias (2008) emphasizes the "striking similarity between the development of … cognitive linguistics … and the recent history of language teaching" (p. 67) and identifies CLT as a usage-based approach.

In fact, Cognitive and Corpus Linguistics have already integrated into CLT, combining the two trends described above. Gries (2008), for example, talks about the "intimate relation" (p. 421) between Corpus and Cognitive Linguistics, and describes their value for language teaching. Meunier (2008) presents an overview of the work in this area and demonstrates a number of similarities between 'the two CLs' and their impact on teaching. Ellis and Cadierno (2009) develop a constructional approach to Second Language Acquisition, using both Cognitive and Corpus Linguistics. Thus, we observe the development of an interdisciplinary approach to language teaching, with Corpus and Cognitive Linguistics as crucial components.

2.2 'The two CLs': implications for this study

Cognitive Linguistics. RFL textbooks contain dozens of example phrases. Some of these share their structure and general meaning and differ only in one or more specific lexical elements (e.g., '**I don't speak [language name]**'). As a result, many structures appear several times throughout the textbook, in which case we call them patterns. In this study, we define a pattern as a *sequence of elements, represented by abstract lexical categories or by specific words, that has a certain structure and meaning and occurs repeatedly in a textbook*. It is important to note that patterns can contain any number of fixed lexical items. This approach is informed by ideas of Cognitive Linguistics, in particular Cognitive Grammar (Langacker, 1987, 2008) and Construction Grammar (Goldberg, 1995; Croft, 2001).

Langacker (1987) introduced the distinction between schematic and specific units, and construed it as a continuum spanning fully specific units (i.e., lexis, expressions and idioms), partially schematic ones (patterns or constructions with one or more open slots) and fully schematic ones (syntactic patterns). Actual units used in conversation are either the result of the reproduction of specific units or of instantiating a partially or fully schematic pattern. That is, speakers produce fully specific entrenched units as well as novel instantiations of the pattern that underlies that specific unit, as both the pattern and some of its instantiations may be entrenched in a speaker's memory, including that of a learner. There is variability, though, in what is better entrenched in learners' memory—the schema or its instantiation(s) (see, e.g., Barðdal, 2008).

Of course, the patterns we extracted from textbooks are not necessarily entrenched patterns for native speakers, let alone for learners, because they were not extracted from actual speech. However, patterns are *potential* constructions in two ways. First, they are provided extensively to learners and there is a consistent finding that learners are sensitive to all kinds of input frequencies (e.g., Ellis, 2002; Ellis & Ferreira-Junior, 2009). Recurrent patterns from a textbook are more likely to be remembered by students—this way extracted patterns from the textbooks may well turn into entrenched schematic patterns in students' idiolects. Second, textbooks ideally describe the ambient language, and many of the recurrent patterns in textbooks are likely to correspond to actual constructions. However, we hypothesize they do not do so to the extent that is desirable, and this assumption is what motivated the present study.

Corpus Linguistics. The implicit assumption behind the comparison of frequencies is that patterns which are more frequent are 'better' in other senses too, e.g. more familiar, more typical, more salient. Although there are conflicting findings on whether frequency necessarily corresponds to prototypicality and cognitive salience (see Gilquin, 2006), we follow Leech (2011) who claims that "if the time wasted teaching rather uncommon structures and weak rules is to be avoided, the 'more frequent = more important to learn' principle should be applied to grammar" (p. 18).

Taking into consideration the theoretical ideas provided above, we proceed to the description of our study.

3. Analyses and results

3.1 Materials

RFL textbooks. We selected three communicatively-oriented breakthrough-level (A1) RFL textbooks that are widely used in modern teaching practice—(I) *Doroga v Rossiyu* (Antonova, Nakhabina, Safronova, & Tolstykh, 2010), (II) *Priglasheniye v Rossiyu* (Korchagina & Stepanova, 2010), and (III) *Zhili-byli* (Miller, Politova, & Rybakova, 2009).

The focus on the breakthrough-level was motivated by two factors. First, textbooks for more advanced levels are based on study materials for previous levels and, therefore, would require additional analysis of patterns already given in those predecessor materials. Second, patterns in breakthrough-level textbooks are given more explicitly, and are therefore easier to determine and extract.

Corpus of spoken Russian language. Since we planned to compare the patterns to the language of everyday communication, we needed a corpus of spoken Russian. The Russian National Corpus appeared to contain only limited amounts of spontaneous everyday speech (approximately 2,000,000 tokens) that was, moreover, unavailable for offline processing due to copyright restrictions, and we were unable to find any other suitable corpus of spoken Russian. Thus, we compiled our own corpus from two available sources that indirectly represent spoken language—dialogs extracted from modern prose, and movie subtitles.

For the first part, we downloaded texts of 20th century Russian prose (1930s—now) from Moshkov Library[1]. From these texts, we extracted all dialogs, resulting in a corpus of 15,443,097 tokens. For the second part, we used Russian movie subtitles (66,150,881 tokens) from the OPUS collection (Tiedemann, 2009). After merging the two sources we obtained a single corpus and automatically annotated it with part-of-speech and lemma information using the TreeTagger tool (Schmid, 1995), because of its reasonably high performance for Russian (Sharoff, Kopotev, Erjavec, Feldman, & Divjak, 2008; Sharoff & Nivre, 2011).

As we already mentioned, the corpus was needed for extracting pattern frequencies. Ideally, this task involves using a syntactically parsed corpus that contains information about phrase structure. Lack of such information causes problems with structurally ambiguous corpus queries—e.g., searching for a pattern '**DET N was found by DET N**' will return both '**The key was found by the door.**' and '**The jewelry was found by the policeman.**', which represent two different

1 http://www.lib.ru. In selecting the prose time frames we followed the existing division of the texts in the library that included the modern prose texts starting from 1930s.

constructions. Although a syntactically parsed corpus allows for a more accurate search, we decided *not* to parse our corpus, since parsing accuracy for Russian in general is not high enough (Sharoff & Nivre, 2011; Nivre, Boguslavsky, & Iomdin, 2008). Considering that we already had a certain error rate because of POS-tagging, additionally including syntactic parsing could lead to unpredictable results in terms of total error rate.

3.2 Procedure

Pattern extraction criteria. From the three textbooks, we extracted all lexico-syntactic patterns that appeared three or more times. We did not consider word-size patterns, i.e., morphology, since in most cases there is only one unmarked way to express a certain meaning by a single morphological unit (e.g., a suffix). We assumed that if a pattern was used in the book only once or twice, it was considered unimportant and could well be omitted by a teacher in the classroom. The choice of three as the threshold is of course rather arbitrary, and there is no evidence that a pattern gets entrenched in the learner's memory after three exposures. On the other hand, we had no way of checking how many times a pattern is used in a classroom. We assume an idealized situation in which teachers strictly follow the textbooks and use recurrent patterns actively in their speech. After a brief review of the textbooks' contents, we decided that three was a reasonable choice. Some examples of extracted patterns are:

(2) *Это твой NOUN(NOM.SG.MASC) ?*
eto tvoy NOUN(NOM.SG.MASC)
this your[NOM.SG.MASC] NOUN(NOM.SG.MASC)
'Is this your NOUN(NOM.SG.MASC)?'

(3) *Вы не знаете, … ?*
vy nye znayete
you(PL) not know:2PL
'Do you know, … ?'

Of course, sometimes the same pattern appeared in each of the three textbooks. However, in most cases such patterns were not completely the same for all three textbooks—the very similar ones still differed in the number of fixed words they contained. For example, while each book introduces phrases that express possession, the respective patterns differ, as shown in Table 1. The degree of specificity for this particular pattern increases from textbook I to textbook III. This does not say much about general differences between the textbooks: for other patterns the opposite ranking was obtained. However, this shows that the possible overlap between different textbooks could not explain the rather low variation between the percentages obtained for each textbook in the results (see Section 3.3).

Table 1. The extracted patterns stating a fact of possession

Book	Pattern				
I	*У*	*PRON(PERS.GEN)/NOUN(GEN)*	*есть*	*NOUN(NOM)*	.
	u	PRON(PERS.GEN)/NOUN(GEN)	yest'	NOUN(NOM)	
	at	PRON(PERS.GEN)/NOUN(GEN)	is	NOUN(NOM)	
	'PRON(PERS)/NOUN have/has a NOUN'				
II	*У*	*PRON(PERS.GEN)*	*есть*	*брат*	.
	u	PRON(PERS.GEN)	yest'	brat	
	at	PRON(PERS.GEN)	is	brother	
	'PRON(PERS) have/has a brother.'				
III	*У*	*меня*	*есть*	*NOUN(NOM.INAN)*[2]	.
	u	menya	yest'	NOUN(NOM.INAN)	
	at	me	is	NOUN(NOM.INAN)	
	'I have a NOUN(INAN).'				

Formulating alternative patterns. For each extracted pattern, we tried to come up with the best alternative pattern with a similar meaning, based on the first author's experience as a Russian native speaker, and preferring patterns we thought were common in everyday speech. Thus, for any given pattern we formulated an *alternative pattern*: a pattern that differed from the original pattern in its form (e.g., the actual elements constituting the pattern, their number and order, the degree of their specificity, etc.), but had a similar meaning. This way we obtained three lists, one for each textbook, of extracted patterns and their respective alternatives.

Corpus querying. We wrote patterns in a format suitable for querying our corpus, and, since the corpus lacked syntactic parsing, post-edited most queries manually. This was needed for processing ambiguous patterns (recall **'DET N was found by DET N'** from section 3.1; see also Gries, 2008). Using the IMS Open Corpus Workbench that provides great capabilities in this respect (Evert & Hardy, 2011), we queried the corpus and extracted the frequency of occurrence of each pattern and three (or fewer, if the total number of matches was lower than three) instances, i.e., actual phrases that matched the pattern.

Native speaker intuitions. Two Russian native speakers judged the patterns, each following a specific procedure. The two procedures complemented each other, since one judge had access only to patterns (containing more general information), while the other worked with the specific examples representing those patterns that were extracted from the corpus. The first judge (the first author) looked through all pattern pairs (the extracted pattern and the suggested alternative) and gave his opinion on which pattern was more typical for everyday spoken Russian. The second judge (a Russian native speaker, an MA student of linguistics) worked only with the actual instances. After removing all pairs in which at least one of the patterns had no match in the corpus, we were

2 Note that the textbook III suggests only inanimate nouns in possessive patterns. Such lack of generalization can sometimes be explained by the limited number of examples in textbooks. However, unless the classroom instruction provides additional examples, learners may have to draw conclusions from the textbook examples only.

left with 319 pairs. Each pair consisted of two groups of three instances of the extracted pattern and three of the alternative, or sometimes fewer than three, for the reasons described above. The judge looked through all pairs and gave her opinion on which group per pair contained phrases that were more typical for everyday spoken Russian. She made her decisions based on what was common for all three instantiations rather than on occasional features of only one example sentence.

Thus, we had three ways to assess the typicality of the extracted patterns: corpus frequency and two types of NS intuitions.

3.3 Results

First, we report on the corpus frequencies of the extracted patterns and their alternatives. In Table 2 we report the numbers of pattern pairs (differentiated per textbook) in which the frequency of the alternative pattern was higher than the frequency of the extracted pattern. There was some variation between the textbooks: for 39 to 46 percent of pattern pairs the alternative had higher frequency in the corpus than the original pattern.

Table 2. Number of more frequent alternative patterns

Book	Total pattern pairs	Pattern pairs in which alternative is more frequent	Percentage of pattern pairs in which alternative is more frequent
I	188	87	46%
II	116	51	44%
III	80	31	39%

The judgment data are provided in Tables 3 and 4, in which we see in how many pattern pairs the alternative pattern (Table 3) or its instances (Table 4) was/were rated as more typical of the everyday language than the extracted originals. These results demonstrate that a substantial number of patterns extracted from the textbooks (from 46 to 53 percent) are considered less typical for everyday spoken Russian than their suggested alternatives.

Table 3. Number of alternative patterns judged to be more typical

Book	Total pattern pairs	Pattern pairs in which alternative is rated higher	Percentage of pattern pairs in which alternative is rated higher
I	188	93	49%
II	116	62	53%
III	80	37	46%

Table 4. Number of alternative instance groups judged to be more typical

Book	Total pattern pairs	Pattern pairs in which alternative is rated higher	Percentage of pattern pairs in which alternative is rated higher
I	158	77	49%

II	90	43	48%
III	71	34	48%

For measuring the inter-rater agreement on the pattern pairs which were present in both judges' lists ($N = 319$), we used Cochran's Q-test[3]. It indicated that differences between the judges were not statistically significant ($Q = 0.82$, $p > .05$).

We finally compared the data for each judge to the pattern frequencies in the corpus. For each pattern pair, we calculated the difference between the frequency of occurrence of the alternative pattern and that of the extracted pattern: $\Delta F = F_A - F_E$. We compared the ΔF values to each judge's data using the point-biserial correlation coefficient (since one of the two variables, namely the judge's opinion, was binary). For the first judge, the correlation was not statistically significant when we considered all the pattern pairs ($N = 384$, $r_{pb} = 0.083$, $p > .05$). However, after removing the six most extreme cases ($|\Delta F| > 30{,}000$) we did obtain a significant correlation for the remaining cases ($N = 378$, $r_{pb} = 0.229$, $p < .001$). The most extreme pattern pairs contributed more than the other ones to the overall values, and if there was a mismatch between the judge's opinion and the corpus evidence for some of these pairs, the results were influenced considerably. For the second judge, as well, the correlation was not significant for all pattern pairs present in her list ($N = 319$, $r_{pb} = 0.02$, $p > .05$), but proved to be significant after we removed the extreme cases ($N = 313$, $r_{pb} = 0.162$, $p < .05$).

Thus, for most pattern pairs the two judges agreed on whether the original pattern or the alternative was more typical for spoken Russian. When comparing the native speaker intuitions to the frequency data, we found that for most pattern pairs the two types of evidence (corpus frequencies and intuitions) were consistent in predicting whether the extracted pattern or the alternative was the more commonly employed option, but not for those cases where there was an extremely large difference between the corpus frequencies of the extracted and alternative patterns.

4. Discussion

Our results show that for 35 to 53 percent of all the recurrent patterns in the textbooks, better alternatives could be found. We used two types of evidence—corpus frequencies and native speaker intuitions. There are reasons to believe that the corpus evidence is more reliable for determining which of the two competing patterns is more typical, since both judges experienced the task as rather difficult. However, despite the view that native speakers are not able to predict which grammatical patterns are more common (Leech, 2011), we found a correlation between the two

3 Since the data provided by the annotators were binary, other metrics such as Cohen's Kappa or Krippendorf's Alpha did not apply.

types of evidence, which allows us to consider the results reliable. The lack of correlation for the extreme cases that we described in the previous section can be explained by a certain degree of subjectivity in our scoring method. This can be demonstrated in the following example (4a-4b):

(4a) *PRON(PERS.DAT)* *надо/нужно* *VB.INF* …
PRON(PERS.DAT) nado/nuzhno VB.INF …
PRON(PERS.DAT) must VB.INF …
'PRON(PERS.DAT) must VB.INF ...'

(4b) *PRON(PERS.NOM)* *должен/должна/должно/должны* *VB.INF* …
PRON(PERS.NOM) dolzhen/dolzhna/dolzhno/dolzhny VB.INF …
PRON(PERS.NOM) have/has to VB.INF …
'PRON(PERS.NOM) have/has to VB.INF ...'

While the original pattern (4a) was judged by both judges to be more typical than (4b), the former occurred substantially less often that the latter. Both patterns (4a) and (4b) can well be used in colloquial Russian, and we believe that judgment data from more native speakers would confirm this.

4.1 Implications for teaching practice

To draw implications for teaching practice, we investigated what the typical shortcomings were of those extracted patterns that received lower scores than their respective alternatives on all three criteria (i.e., lower corpus frequency and judged to be less typical than the alternative by both judges). There were three general groups of shortcomings: redundancy, non-prototypicality in formulaic sequences and non-prototypical word order.

Redundancy. Low-score patterns often contain unnecessary elements that are normally omitted in colloquial language (recall examples 1a-1b), as in the following examples.

- *Pronouns:*

(5a) *Я* *думаю,* *что* …
ya dumayu chto
I think that
'I think that ...'

The alternative (5b) eliminates the personal pronoun and the conjunction:

(5b) *Думаю,* …
dumayu
think:1S
'I think ...'

- *Verbs:*

(6a) *Куда* *PRON(PERS.NOM.2)* *идёшь* *?*
kuda PRON(PERS.NOM.2) idyosh
where you go
'Where are you going?'

The alternative (6b) leaves out the verb:

(6b) *PRON(PERS.NOM.2) куда ?*
PRON(PERS.NOM.2) kuda
you where
'Where are you going?'

- *Adverbs:*

(7a) *Сколько сейчас времени ?*
skol'ko seychas vryemyeni
how.much now time:GEN
'What time is it now?'

Even the colloquial English translation of (7b) omits the adverb:

(7b) *Сколько времени ?*
skol'ko vryemyeni
how.much time:GEN
'What time is it?'

Additionally, we discovered a group of patterns that were extracted from example dialogs in the textbooks and, thus, should be characterized by ellipsis, but fail to do so. We already mentioned in the introduction that colloquial language contains a lot of linguistically incomplete expressions. We believe that examples like (8a-8b) demonstrate the value of colloquial language for CLT at the breakthrough level:

(8a) *– Вам нравится/нравятся этот/эта/это/эти NOUN(NOM) ?*
– Да, мне нравится/нравятся этот/эта/это/эти NOUN(NOM).
vam nravitsya/nravyatsya etot/eta/eto/eti NOUN(NOM)
da mnye nravitsya/nravyatsya etot/eta/eto/eti NOUN(NOM)
you(S.DAT) please:3.PRES(REFL) this/these NOUN(NOM)
yes I(DAT) please:3.PRES(REFL) this/these NOUN(NOM)
'– Do you like this/these NOUN? – Yes, I like this/these NOUN.'

A more concise and natural answer is given in (8b):

(8b) *– Вам нравится/нравятся этот/эта/это/эти NOUN(NOM) ?*
– Да, нравится/нравятся.
vam nravitsya/nravyatsya etot/eta/eto/eti NOUN(NOM)
da nravitsya/nravyatsya
you(S.DAT) please:3.PRES(REFL) this/these NOUN(NOM)
yes please:3.PRES(REFL)
'– Do you like this/these NOUN? – Yes, I do.'

Use of non-prototypical sequences. Native speakers often have a clear preference for a certain sequence of words over another one (e.g., Wray, 2000). Although this should ideally be reflected in the textbooks' content, there were a number of cases where the extracted pattern contained the less common phrase. Some of these included politeness expressions:

(9a) *Скажите, пожалуйста, где ... ?*
skazhitye pozhaluysta gdye
tell:2PL please where

'Tell me please, where … ?'

The alternative that scored better was:

(9b) *Вы* *не* *подскажете,* *где* ... *?*
vy nye podskazhetye gdye
you:PL not tell:2PL.FUT where
'Could you tell me please, where … ?'

The two patterns in (9a-9b) are stylistically equal, so the preference can be explained only by the conventions of language use. However, in some other pairs the pattern and the alternative index different styles. The patterns in (10a) and (11a) are more formal than the alternatives in (10b) and (11b), respectively:

(10a) *Приглашаю* *PRON(2)* *в* *гости* .
priglashayu PRON(2) v gosti
invite:1S PRON(2) to guests
'I'd like to invite you for a visit.'

(10b) *Заходи/заходите* *в* *гости* .
zahodi/zahoditye v gosti
call.on:IMP to guests
'Come and see us sometime.'

(11a) *Какой/какая/какое/какие* *это* *NOUN(NOM)* *?*
kakoy/kakaya/kakoye/kakiye eto NOUN(NOM)
what this NOUN(NOM)
'What is this/these NOUN? / How does this/these NOUN look?'

(11b) *Что* *это* *за* *NOUN(NOM)* *?*
chto eto za NOUN(NOM)
what.kind this what.kind NOUN(NOM)
'What kind of NOUN is this?'

In the next pair, (12a) can only be used in a very narrow range of contexts because of its specific meaning, while the alternative (12b) with the verb in the past tense (rather than in the present tense) appears to be more generally useful.

(12a) *Что* *говорит* *NOUN(NOM)* *?*
chto govorit NOUN(NOM)
what say:3S NOUN(NOM)
'What does NOUN say?'

(12b) *Что* *сказал/сказала* *NOUN(NOM)* *?*
chto skazal/skazala NOUN(NOM)
what say:PAST NOUN(NOM)
'What did NOUN say?'

However, this last example may be controversial, because the meanings of (12a) and (12b) are slightly different. The extracted pattern (12a) would more often be used for asking about somebody's opinion (synonymous to '**What does he think of it?**'), while the alternative (12b) suits a context in which someone is asking about past events (as in '**And what did father say after you**

broke the glass?'). These examples illustrate a general problem for our approach: it is widely assumed that absolute synonyms do not exist in general, so we should always be alert to the fact that the two Russian patterns may not truly be alternative realizations of the same content. However, the arguments of frequency and usefulness extend to communicative contexts and stylistic preferences.

These examples raise the question which register should be taught to beginners. Although in general communicative competence presupposes the ability to engage in informal as well as formal communication (e.g., Richards, 2006), most students at the breakthrough level will more often use their Russian in informal everyday situations, as we mentioned in Section 1. This is also implied by the standards of the Common European Framework of Reference for Languages. These arguments support our contention that it is the informal register that should be preferred for general usage textbooks.

Word order. We found a small number of patterns with a non-prototypical word order, e.g.:

(13a)	*Где*	*ты*	*VB(2S.PST.IPFV)*	*раньше?*
	gdye	ty	VB(2S.PST.IPFV)	ran'she
	where	you(SG)	VB(2S.PST.IPFV)	before
	'Where did you VB before? / Where have you PP before?'			

In (13a) the word order is slightly changed from the more acceptable and frequent alternative in spoken Russian given in (13b):

(13b)	*Где*	*ты*	*раньше*	*VB(2S.PST.IPFV)*	*?*
	gdye	ty	ran'she	VB(2S.PST.IPFV)	
	where	you(SG)	before	VB(2S.PST.IPFV)	
	'Where did you VB before? / Where have you PP before?'				

4.2 General conclusion

Our study is intended to help textbook writers to avoid some typical shortcomings when deciding which content to include. It demonstrates how Cognitive and Corpus Linguistics can contribute to textbook writing. Most research in this respect has been done for English, while for Russian this is the first study of its kind. While most previous studies have focused on words and their frequencies, we have analyzed patterns, or partially schematic constructions. Our methodology represents a general shift towards Cognitive Linguistics as a useful theoretical framework for Second Language Acquisition, and reflects our confidence in the value of the usage-based approach as the basis of a descriptively adequate theory of language, and, by extension, of sound teaching.

In conclusion, we note a couple of suggestions for future study. First, the alternative patterns were suggested by only one native speaker, so they should be reconsidered in further research by a larger group of native speakers. Second, a number of improvements can be made to the corpus.

More accurate results will be obtained if larger amounts of syntactically parsed and semantically annotated spoken language data become available.

References

Antonova, V. E., Nakhabina, M. M., Safronova, M. V., & Tolstykh, A. A. (2010). *Doroga v Rossiyu: uchebnik russkogo yazyka (elementarnyy uroven)* [The Way to Russia: Russian Language Textbook (Elementary Level)] (6th ed.). Moscow: MSU.

Barbieri, F., & Eckhardt, S. E. B. **(**2007). Applying corpus-based findings to form-focused instruction: The case of reported speech. *Language Teaching Research*, *11*(3), 319-346.

Barðdal, J. (2008). *Productivity: Evidence from Case and Argument Structure in Icelandic.* Amsterdam: John Benjamins.

Berns, M. (1990). *Contexts of Competence: Social and Cultural Considerations in Communicative Language Teaching*. New York: Plenum Press.

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, *24*(2), 199-208.

Broccias, C. (2008). Cognitive linguistic theories of grammar and grammar teaching. In S. De Knop, & T. De Rycker (Eds.), *Cognitive Approaches to Pedagogical Grammar* (pp. 67-90). Berlin: Mouton de Gruyter.

Carter, R., Hughes, R., & McCarthy, M. (2000). *Exploring Grammar in Context: Upper-intermediate and Advanced*. Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (1994). Grammar and the spoken language. *Applied Linguistics*, *16*(2), 141-157.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C., & Cadierno, T. (2009). Constructing a second language: Introduction to the special section. *Annual Review of Cognitive Linguistics*, 7, 111-139.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7, 188-221.

Ellis, N. C., & Robinson, P. (2008). An introduction to Cognitive Linguistics, Second Language Acquisition, and language instruction. In P. Robinson, & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 3-24). New York: Routledge.

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference, University of Birmingham, UK*.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1996). *Grammar Patterns 1: Verbs*. London: HarperCollins Publishers Ltd.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language teaching. *Language Teaching*, *40*(2), 97-118.

Gilquin, G. (2006). The place of prototypicality in corpus linguistics. Causation in the hot seat. In S. T. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (pp. 159-191). Berlin: Mouton de Gruyter.

Glisan, E. W., & Drescher, V. (1993). Textbook grammar: does it reflect native speaker speech? *The Modern Language Journal*, 77*(*1), 23-33.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gries, S. T. (2008). Corpus-based methods in analyses of Second Language Acquisition data. In P. Robinson, & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 406-131). New York: Routledge.

Hornby, A. S. (1954). *Guide to Patterns and Usage in English* (2nd ed.). London: Oxford University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Hutchinson, T., & Waters, A. (1987). English for Specific Purposes: A learning-centered approach. Cambridge: Cambridge University Press.

Johnson, K., & Johnson, H. (1998). Communicative methodology. In K. Johnson and H. Johnson (Eds.), *Encylopedic Dictionary of Applied Linguistics*. Oxford: Blackwell. 68-73.

Korchagina, E. L., & Stepanova, E. M. (2010). *Priglasheniye v Rossiyu. Chast I. Elementarnyy prakticheskiy kurs russkogo yazyka. Uchebnik* [Invitation to Russia. Part 1. Elementary Practical Course of Russian Language. Textbook] (6th ed.). Moscow: Russkiy yazyk.

Langacker, R. W. (1987). *Foundations* of C*ognitive Grammar*: *Theoretical Prerequisites* (*Vol*. 1). Stanford, CA: Stanford University Press.

Langacker, R. W. (1999). *Grammar and Conceptualization*. Berlin: Mouton de Gruyter.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Leech, G. N. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A Taste for Corpora* (pp. 7-32). Amsterdam: John Benjamins.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

McCarthy, M. (2004). *Touchstone: From Corpus to Course Book*. Cambridge: Cambridge University Press.

Meunier, F. (2008). Corpora, cognition and pedagogical grammars: An account of convergences and divergences. In S. De Knop, & T. De Rycker (Eds.), *Cognitive Approaches to Pedagogical Grammar* (pp. 91-119). Berlin: Mouton de Gruyter.

Miller, L. V., Politova, I. V., & Rybakova, I. Y. (2009). Zhyli-Byli… 28 urokov russkogo yazyka dlya nachinayuschikh: Uchebnik [Once upon a Time… 28 Russian Lessons for Beginners: Textbook]. Saint-Petersburg: Zlatoust.

Nivre, J., Boguslavsky, I. M., & Iomdin, L. T. **(**2008). Parsing the SynTagRus Treebank of Russian. In D. Scott, & H. Uszkoreit (Eds.), *22nd International Conference on Computational Linguistics. Proceedings of the Conference* (pp. 641-648). Manchester: Association for Computational Linguistics.

Richards, J. C. (2006). *Communicative Language Teaching Today*. Cambridge: Cambridge University Press.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge: Cambridge University Press.

Savignon, S. J. (2002). Communicative language teaching: Linguistics theory and classroom practice. In S. J. Savignon (Ed.), *Interpreting Communicative Language Teaching: Contexts and Concerns in Teacher Education* (pp. 1-27). New Haven: Yale University Press.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In S. Armstrong, & E. Tzoukermann (Eds.): *Proceedings of the EACL SIGDAT-Workshop* (pp. 47-50). Association for Computational Linguistics.

Schönefeld, D. (2006). Constructions. *Constructions, Special Volume 1-1/2006*, retrieved on 1 April 2012, from http://elanguage.net/journals/constructions/article/download/16/36

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., & Divjak, D. (2008). Designing and evaluating Russian tagsets. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association.

Sharoff, S., & Nivre. J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Kompyuternaya lingvistika i intellectualnye tekhnologii: po materialam yezhegodnoy konferencii "Dialog" (Bekasovo, 25-29 maya 2011 g.) [Computational Linguistics and Intellectual Technologies: Based on Materials from Annual "Dialog" Conference (Bekasovo, May 25-29, 2011)]*. Moscow: RGGU.

Spada, N. (2007). Communicative language teaching: Current status and future prospects. In J. Cummins, & C. Davis (Eds.), *Kluwer Handbook of English Language Teaching* (pp. 271-188). Amsterdam: Kluwer Publications.

Tiedemann, J. (2009). News from OPUS—A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007* (pp.237-248). Amsterdam: John Benjamins.

Willis, D. (1990). *The Lexical Syllabus: A New Approach to Language Teaching*. London: Collins ELT.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463-489.