Russian Language Journal

Volume 63 | Issue 1 Article 11

2013

Using a Corpus-Based Approach to Russian as a Foreign Language Materials Development

Edie Furniss

Using a Corpus-Based Approach to Russian as a Foreign Language Materials Development

Edie Furniss

Introduction

The increase in availability and sophistication of corpora in recent years has facilitated the application of usage-based approaches to language pedagogy. Although the use of corpus data is certainly not without its difficulties, it offers great pedagogical promise. Corpora, consisting of natural language culled from a multitude of sources and genres, provide valuable information about language in use. While a corpus can provide us with contextualized linguistic data and statistics on the behavior of lexicon (with respect to frequency and collocation), a connection needs to be forged between the data and their practical use. Two main areas ripe for the application of corpus linguistics are data-driven learning and materials development. Data-driven learning concerns the study of language by learners who use corpora to obtain raw data for analysis (see, for instance, Johns 1991, Gavioli & Aston 2001, Varley 2009).

The focus of this paper, however, is the practical application of corpus data to the development of foreign language teaching materials, specifically for the Russian language learner audience. A corpus-based approach, in my view, can enable the creation of textbooks that better serve their users, as it ensures that the language presented is contextualized and reflective of actual usage. In this paper, I will discuss the arguments for using corpora to inform pedagogical materials, how ESL/EFL textbooks have implemented a corpus-based approach, and practical guidelines for employing such an approach with Russian language materials (with reference to vocabulary selection, potential exercises and activities, and learning context).

Conrad (2000) anticipated three potential shifts in grammar instruction as a result of corpus-based linguistic research: the replacement of large and comprehensive English grammars with smaller, register-specific ones; the combination of grammar and vocabulary; and the move towards a greater focus on appropriateness of use, rather than structural accuracy (p. 549). While these changes are increasingly being

realized in materials for English language learners, the field of Russian as a foreign language (RFL) pedagogy has yet to see a similar response to budding corpus-based research. Regarding the first shift, a comprehensive learner's dictionary of Russian—an analogue to the corpus-based English-language Collins COBUILD series—has yet to be developed. And, most crucially, there is a need for commercially available RFL grammars and textbooks informed by corpus data. A focus on the combination of grammar and vocabulary, or lexicogrammar, is notably absent from RFL materials, which tend to approach these aspects of language as two discrete systems. Lexicogrammar is defined as "the lexicon and grammar of a language, taken together as an integrated system" (Halliday, Teubert, Yallop, & Čermáková, 2004, p. 169). When performing corpus analyses, researchers must note the morphological and syntactic restraints of lexical items and phrases, paying attention to how grammar and vocabulary inform one another. Finally, emphasis on appropriateness of use is needed in materials, meaning that they must raise awareness of context, and each context's corresponding pragmatic guidelines. However, structural accuracy should not be sacrificed in service of this goal. Because of the greater morphological complexity of the Russian language (as compared with English), an explicit focus on grammar is essential, but should be combined with contextual corpus-driven analysis and engagement.

Perhaps the greatest issue in the current state of RFL materials that can be addressed by corpus linguistics is the need for a systematic approach to the Russian language in use, with reference to frequency and actual linguistic behavior. O'Keefe, McCarthy, and Carter (2007) noted that "numerous studies have shown us that the language presented in textbooks is frequently still based on intuitions about how we use language, rather than actual evidence of use" (p. 21). Corpus data can be used to compare language in use with the lexicogrammatical elements featured in textbooks and other materials. Because corpora are an excellent source of frequency information, authors of instructional materials can harness their power to better select and accurately prioritize the language presented. This kind of investigation requires case studies focusing on particular linguistic elements. Conrad (2004) examined four ESL textbooks in order to compare the treatment of *though* with corpus data on its frequency and details of its usage. She found that while linking adverbials were included in the textbooks, *though*, the most frequently used linking adverbial, was only included in one of the four

textbooks, and was there covered incompletely (only the contrastive, not concessive use was mentioned). Flowerdew (1998) conducted a similar comparison of academic writing textbooks and corpus data, finding that cause/effect markers in English, which are commonly used in a corpus of academic English, are inadequately covered in English for academic purposes (EAP) textbooks. Omissions of this sort can confuse learners by promoting an inaccurate picture of language usage, resulting in production of language that is "often stilted, too formal and too high-level; and when it is analysed it is seen that the most common words are used less frequently, and in fewer contexts, than they would be by native speakers of English" (Tomlinson, 1998, p. 27). Corpora are continually being developed and expanded, providing a growing body of data on actual language use. It seems only logical to use them to create materials that will better inform language learners and assist them in becoming more fluent users of the target language.

However, real-life language from a corpus can be messy and difficult to analyze, and generally does not lend itself to succinct usage explanations, like those found in traditional grammars. Conrad (2004) addressed the reluctance that many teachers feel when responding to a student's usage query with the answer 'It depends': "With analyses [comparing corpus data with textbooks], we find out not only that the answer to most questions about language use is 'it depends,' but we can also answer the question 'What does it depend on?'" (p. 80). In order to answer that question, a highly nuanced examination of the data is required, as well as a reevaluation of prior conceptions of lexicon and grammar, on the parts of both teacher and students.

Using corpus data to answer the question "What does it depend on?" supports Larsen-Freeman's (2003) concept of *grammaring*—"the ability to use grammar structures accurately, meaningfully, and appropriately" (p. 143). With examples from and statistics on real-world usage, instructors and materials writers can better define what constitutes accuracy, meaningfulness, and appropriateness not only on the grammatical level, but on the lexical level as well. This requires attention to the forms themselves, as well as the contexts in which they appear. Such an approach can assist in the development of a genre-based syllabus, as in the case of Chang and Kuo (2011), who combined corpus and genre analysis in preparing online materials for an EAP course. The authors conducted a genre analysis of the texts in a corpus of research articles, then a text analysis focusing on lexicogrammatical elements,

which led to the creation of PDFs of the texts tagged with each rhetorical move and accompanied by notes on linguistic features. Traditionally, language teaching materials have been focused on the sentence level, which naturally leads to decontextualized linguistic examples that are displaced from the world of real language use.

Discursive analysis of language is clearly necessary if materials writers want to accurately reflect authentic usage, and should be combined with lexicogrammatical analysis to that end as well. This means attention not only to frequency and use of individual lexical units, but also to a particular area of language that is all too often underrepresented in instructional materials: formulaic sequences. Wray (2000) defines a formulaic sequence as:

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (p. 465)

Corpora provide data that textbook writers and instructors can use in the identification and description of these sequences, seeing as they contain natural language data that is representative of real language use (Boers & Lindstromberg, 2009). Such information can be obtained to create lexicogrammatical profiles consisting of collocates, chunks/idioms, syntactic restrictions, semantic restrictions, prosody, and other relevant or recurring features (O'Keefe et al., 2007). These profiles would be more informative to language learners than the vocabulary lists consisting of single words and their English equivalents that pervade RFL materials.

The question, however, arises: which formulaic sequences should be taught? Boers and Lindstromberg (2009) propose that in making such decisions, learning context and learner needs are paramount: social routine formulae may be most useful for students in naturalistic environments, discourse organizers for students of academic writing, and referential language for learners in traditional foreign language classroom settings. American learners of Russian in the U.S. generally study the language in preparation for study abroad and/or to engage with Russian literature in the original. Thus, formulaic sequences found in corpora of informal conversation, university lectures, and literature (particularly of the 19th century) should be emphasized in instructional materials.

One of the most comprehensive and accessible corpora of the Russian language is the online Russian National Corpus (http://www.ruscorpora.ru/en); also see the large Russian corpora at IntelliText (http://smlc09.leeds.ac.uk/itb) and Sketch Engine (https://the.sketchengine.co.uk/auth/corpora/). The Russian National Corpus, containing over 150 million words in texts from the mid-18th to early 21st centuries, is the work of linguists from around Russia and is funded by Russian Federation governmental grants. It consists of a general corpus as well as the following subcorpora: syntactic (with in-depth annotations); mass media (containing current texts from newspapers); parallel Russian-English (to facilitate comparison of translations between the two languages); educational (developed for Russian elementary and secondary schools); dialectical (featuring various varieties of spoken Russian); poetry (with search parameters like meter and rhyme type); spoken (including public and private speech, and movie dialogue from 1930 on); accent (focusing on word stress); and multimedia (accompanied by video clips featuring the queried word or phrase). Additionally, it is possible to tailor the corpus data to one's needs by creating custom-made subcorpora. One can personalize the subcorpus by deciding what author(s), texts, speaker/author gender(s), year(s) of publication, text genre(s) and type(s), text setting(s), and subject matter to include. Search results can then be downloaded and manipulated using Microsoft Excel. The Russian National Corpus is obviously an extremely useful and flexible resource that should be an essential tool in the development of language teaching materials, but has not yet been used extensively for that purpose.

Given the great variety of texts included in the corpus, there is a multitude of potential applications of the corpus data to pedagogy. For example, the mass media subcorpus could be used to determine the most frequent collocates of a key word for inclusion in political Russian materials, such as *povyshenie* (e.g., …*nalogov*, *tarifov*, *zarabotnoi platy*, etc.). Lists of uses of difficult-to-translate words such as the article *ved'* or the adjective *sploshnoi* can be easily compiled, in order to study their behavior in discourse. The texts found in the parallel corpus could provide aspiring translators with concrete data on typical translations of a queried word or phrase, which could then be analyzed, compared, and critiqued. Close synonyms, like *druzheliubnii, druzheskii, druzhestvennii*, and *druzhnii*, could be disambiguated by examining the subtleties of their usage in authentic contexts, across different genres.

The only major criticism of the Russian National corpus is the fact that its source texts cannot be read in full via its online interface; rather, search queries are returned with only the immediate context (from one to two sentences to one paragraph). However, users may obtain a significantly reduced (consisting of about one million words) offline version of the corpus by signing and submitting a license agreement. Of course, the complete context (including detailed information on participants, and non-verbal elements of language) may arguably *never* be available, although corpus excerpts can be expanded by integrating audio (to highlight pronunciation and prosody) and video (to include information on gesture, gaze, and so on). So far, only the multimedia subcorpus allows for this possibility, but contextualized examples of usage can be located on the Internet (such as authentic video available on YouTube) and used to support data from the corpus.

Usage patterns can still be effectively extracted from corpus data. For example, the phrase *nichego sebe* [wow] appears 35 times in the nonpublic spoken subcorpus of the Russian National Corpus, as a stand-alone exclamation (23 times), often in response to an interlocutor, and sometimes modifying another word (ten times). In two instances, the function of the phrase was ambiguous (due to lack of punctuation). As an exclamation, *nichego sebe* indicates surprise or disbelief, as in the following conversation between a 69-year-old woman (Speaker 1) and a 45-year-old woman (Speaker 2):

> Speaker 1: *On znachit e… m… Ia khochu, govorit, vas / priglasit' na tusovku. (so smekhom v golose) Na kakuiu tusovku?* [He, then… um… I want to, he says, invite you to a party. (with laughter in her voice) What kind of party?]
>
> Speaker 2: ***Nichego sebe**! (smeetsia)* [**Wow**! (laughs)]
>
> Speaker 1: *A on govorit, vot v Ostankino / est' dlia veteranov.* [And he says, in Ostankino there's one for veterans.]
>
> Source: Tea-table talk // M.V. Kitaigorodskaia, N.N. Rozanova. Muscovites' speech: Communicative-culturological aspect. Moscow, 1999, 1985-1992.

In contrast, in this following conversation between two 18-year-old females looking at a photograph, the use of *nichego sebe* similarly displays surprise, but instead modifies a noun rather than standing alone as an interjection:

> Speaker 1: *Takaia priam*… [A real…]
>
> Speaker 2: *Rokovaia*… [vampy…]

Speaker 1: *devochka*… [little girl…]
Speaker 2: ***Nichego sebe*** *devochka… Chto zh devochka-to? Takaia devushka uzhe / vzroslaia…* [**Quite** a little girl… Why a little girl? A young lady already / grown-up]
Speaker 1: *Nu devushka / ladno…* [Fine, a young lady, okay…]
Source: Looking at photos // From Ulianovsk University materials, 2006

While the full conversations are not available from the Russian National Corpus, these excerpts still provide corpus users with valuable information. Such data can be used to raise awareness of the use of *nichego sebe* in speech by showing that it is more commonly used as an interjection and a comment on an interlocutor's utterance than as a modifier. Further, a materials writer could examine the corpus for other interjections used as comments in order to find which phrases are most commonly used, in what contexts, and how their meanings and tone may differ (is a particular interjection positive or negative? and so on).

This lack of full text accessibility in the Russian National Corpus underscores the relevance of two problematic issues in the application of corpus data to pedagogy, as noted by Flowerdew (2009): an emphasis on bottom-up processing of text; and decontextualized (and therefore not transferable to pedagogy) data. This echoes Widdowson's (2000) concern that such data are inauthentic, being stripped of sufficient context (including the perspectives of the participants). Flowerdew's (2009) first criticism is leveled at the practice of heavy reliance on concordance lines, which may not provide sufficient context. In terms of the published corpus-based pedagogical materials reviewed later, however, concordance lines are rarely used. Instead, lexicogrammatical features are presented in discourse, or separately with commentary on usage. Presumably, this is done so that corpus-based textbooks closely resemble traditional materials—a concordance looks foreign and might discourage users who find it too technical.

Flowerdew (2009) noted two more issues with using corpus-based methods: the prominence of the inductive approach in corpus-based pedagogy, and the difficulty of choosing the appropriate corpus. While traditionally an inductive approach has been preferred, it is by no means necessary. A mix of both inductive and deductive activities can be created on the basis of corpus data. Further, the corpus data can be presented in a subtle, not overly technical way (Conrad, 2000). Regarding Flowerdew's (2009) final criticism, the problem of limited corpus

availability, including specialized corpora, is slowly being remedied, particularly for English. Increased corpus diversity will simplify the task of choosing an appropriate corpus.

While these are valid arguments, still Mauranen (2004) contends that "corpus data is light years ahead of invented examples in authenticity, and to make the most use of that data is a matter of pedagogic intervention in the learning process" (p. 94). These 'invented examples' are ubiquitous in RFL instructional materials. One notable exception is Rifkin's (1996) *Grammatika v Kontekste: Russian Grammar in Literary Contexts*, which uses authentic literary and journalistic texts to contextualize language. A focus on literary language may be worthwhile, depending on the context and learner needs; many students in college-level Russian programs will enroll in literature survey courses (most commonly with an emphasis on the classics of the 19th century). The potential of reading Russian prose and poetry in the original has always been a strong motivating factor for generations of Russian language learners. Regrettably, many of these learners are ill-prepared for reading Russian literature, due to the lack of targeted instruction in Russian literary discourse. Rifkin (1996) has presented readers with excerpts of this sort, but a more systematic approach to the material could maximize student learning potential. This could be accomplished through the use of a corpus consisting of literary texts of the era in question, which could be created, as mentioned earlier, using the personalized subcorpus tool on the Russian National Corpus website. Such a corpus of Russian literature could be explored with reference to word and formulaic sequence frequency in order to better inform teaching materials.

Opportunities exist for the adoption of corpus-informed approaches to materials development, but instructors and authors need to be made aware of them. Additionally, Reinhardt (2010) proposes that, in order for corpus linguistics to have a bigger effect on language pedagogy, it is necessary to have available "corpus-informed materials, corpus analysis tools, and well-designed corpora [that] are simply more numerous, accessible and user friendly, and preferably web-based" (p. 246). While corpus-informed EFL materials are gradually filling this need, the same is not true for RFL. One reason for the deficit is the much lower demand for Russian linguistic materials, which means less incentive for developing sophisticated corpus-based RFL textbooks and specialized Russian-language corpora. Additionally, corpora of English (and, consequently, corpus tools created specifically for the English

alphabet and English morphology and syntax) have been in existence much longer than corpora of other languages. Now that corpora and corpus tools are becoming more widely available for Russian, a shift in the conceptualization of language usage and pedagogy is necessary if corpus-informed approaches are to flourish in the Russian language teaching community.

Corpus-based Materials

In their guide for creating materials using data from MICASE (Michigan Corpus of Academic Spoken English), Simpson-Vlach and Leicher (2006) advocate a combination of discourse- and corpus-based approaches, which, in union, can "more easily guide our students in learning pragmatically and sociolinguistically *likely* and *appropriate* uses of language, rather than just grammatically correct uses" (p. 267). This point is crucial in relation to materials development, as it requires examining language data holistically and in context. In order to achieve this end, the authors present materials based on a transcript of language in use drawn from the corpus, accompanied by notes, discussion questions, and exercises. Worksheets of this type can be developed by instructors without too much difficulty, as Simpson-Vlach and Leicher (2006) suggest, by targeting situational, functional, or pragmatic language usage, or by focusing on specific lexicogrammatical features.

Using a complete text (in this case, a transcript of a spoken interaction) as the unit of analysis is one way to integrate corpus data into instructional materials. In *Exploring Spoken English*, Carter and McCarthy (1997) use this approach, providing authentic spoken texts from the CANCODE (Cambridge-Nottingham Corpus of Discourse in English) belonging to several spheres, or genres, of interaction: narrative; identifying; language-in-action; comment-elaboration; service encounters; debate and argument; language, learning and interaction; decision-making/negotiating outcomes. Detailed annotations containing linguistic and cultural observations follow each text, and recordings of the texts themselves can be accessed via an accompanying cassette tape. Given the absence of exercises, *Exploring Spoken English* is intended as an awareness-raising tool. Such materials are useful for learners, as they are all too infrequently exposed to annotated natural discourse. Since the genres presented in *Exploring Spoken English* are universal, a similar schema could be used in the development of comparable Russian language materials. However, a more effective method might be the use

of shorter texts, particularly for beginning and intermediate learners, seeing as a single text can often be long, disjointed, and unwieldy.

Real Grammar, a corpus-based textbook for learners of English, by Conrad and Biber (2009), consists of 50 units addressing various lexicogrammatical topics. The authors consistently make distinctions throughout the text between the lexicogrammar of speech and the lexicogrammar of writing. In the front matter, Conrad and Biber (2009) describe their methodology in regards to adapting corpus texts for presentation in the textbook: the replacement of difficult vocabulary with easier vocabulary; simplification of long and complex sentences; revision of academic writing discourse excerpts; removal of some fillers and false starts; and standardization of syntax through the addition of punctuation. Their rationale for these modifications is that "it is important for the language of the corpus extracts not to overwhelm students or to take their attention away from the structure that is being practiced" (Conrad & Biber, 2009, p. ix). Adjustments of this nature are necessary in the adaptation of authentic language data. One of the sentiments expressed in the introductions to many of the textbooks discussed here was the desire to produce explanations and activities that are similar to those of existing materials, resulting in a more user-friendly product. After all, the purpose in using corpus data is to give learners a clear look at real-life language usage and to motivate them in their language study, not to overwhelm them. Instructors and materials writers, then, must use their judgment in deciding what revisions, if any, are necessary.

McCarthy, O'Dell, and Shaw's *Vocabulary in Use* (1997) is composed of one hundred vocabulary units, the selection and content of which were informed by the Cambridge International Corpus. McCarthy (2004) described the method for vocabulary selection for the *Vocabulary in Use* series: comparison of the vocabulary lists to corpus data to ensure relevance; identification of a basic vocabulary for beginning learners; and inclusion of lexicon in the most frequent (according to corpora) contexts and situations. The units each span two pages and are organized by topic—grammatical (e.g., uncountable nouns); functional (e.g., everyday problems); metaphorical (e.g., idioms describing feelings and mood); thematic (e.g., sports); and others (e.g., the language of signs and notices; discourse markers). There are no textbooks, to my knowledge, that address Russian vocabulary in a similar way. *Vocabulary in Use* is focused not simply on discrete lexical items, but on lexicogrammar, including

formulaic sequences. Lexicogrammatical features are given in context (generally one to two sentences, sometimes in longer discourse) with commentary on semantics and usage, and then followed by a variety of exercises (cloze; rewording; matching question and response; open-ended sentence completion; picture labeling, and more).

Exploring Grammar in Context, by Carter, Hughes, and McCarthy (2000), also uses the CANCODE corpus to inform selection of grammatical structures and to obtain authentic language excerpts. Units are ordered by the following categories: tenses; modals; choosing structures; around the noun; and exploring spoken grammar in context. The textbook provides a wide variety of exercises to engage with the material, including: identifying targeted grammatical structures in context; making observations about usage based on the texts; rewriting sentences to use target structures; selecting the appropriate structure depending on the context (e.g., deciding between the use of would/will on p. 41); raising awareness of collocations and fixed expressions with notes on usage; working inductively with grammar in context to draw conclusions about usage; engaging with authentic language outside of class (e.g., "Look at an editorial in an English newspaper, or any other text where someone is presenting arguments or opinions, and note how *it, this* and *that* are used to refer to the points the writer is making" p. 92); ordering conversation turns; matching expressions with their meanings; and other common activities (error correction, cloze, etc.). Like the other titles reviewed here, *Exploring Grammar in Context* contextualizes language usage and provides students with exercises and examples that facilitate the acquisition of the most necessary grammatical structures.

Meanings and Metaphors (Lazar, 2003), a textbook for intermediate learners of English, introduces high-frequency figurative language, as determined by corpus analysis. Chapters are organized by topic and genre: parts of the body; weather; plants; colors; poems; proverbs; and so on. *Meanings and Metaphors* provides activities that raise awareness of metaphor, contextualize idiomatic phrases in discourse (including authentic texts from advertising), and ask students to identify metaphors independently. O'Keefe, McCarthy, and Carter (2007) discussed the need to consider metaphor in vocabulary for language learners, stating that "much advanced level vocabulary pedagogy will be concerned with dealing with less frequent, extended and metaphorical sense of words" (p. 51). Corpora can make identifying such figurative language easier, while providing information on context and frequency.

Touchstone is a corpus-informed sequence of materials by McCarthy, McCarten, and Sandiford (2005). The authors, in undertaking the corpus research necessary for the series, named the following goals: "to identify authentic, motivating language; to weave [the] findings into a carefully crafted syllabus; to create course books that are familiar in structure and easy to use" (McCarthy, 2004, p. 15). Level 2 of the *Touchstone* series (for high beginners) is divided into twelve units, each focusing on a different topic (e.g., health, growing up, at home, etc.). Each unit contains the following sections: function/topic; grammar; vocabulary; conversation strategies; pronunciation; listening; reading; writing; vocabulary notebook; and free talk. The conversations included in *Touchstone* are constructed by the authors, but with the corpus data as a guide, and the textbook is accompanied by an audio CD that reproduces these dialogues. This strategy may be most appropriate for beginners' level materials like *Touchstone*, as truly authentic corpus excerpts may be inaccessible to learners with limited lexicogrammatical knowledge. Indeed, Reinhardt (2010) suggested that, in textbooks for beginning learners, "dialogues can reflect corpus-based findings, vocabulary can be presented with collocational and frequency information, and grammatical explanations can be contextualized in discussions of register appropriateness" (p. 247).

The focus on conversation strategies (based, of course, on corpus examples) is noteworthy; the authors describe a variety of everyday issues in pragmatics (responding to suggestions, using *I mean* to correct yourself when you say the wrong word or name, agreeing to something with *All right* and *OK*, and so on), providing lists of common expressions to use according to the situation and opportunities to apply the given information in matching and cloze (fill-in-the-blank) exercises, discourse completion tasks, and role plays. For example, on page 71:

> ***Strategy plus*** *I guess*
>
> You can use ***I guess*** when you're not 100% sure about something, or if you don't want to sound 100% sure.
>
> *– I guess I need to keep this job.*
>
> *– Yeah, me too, I guess.*

This type of specific, focused instruction on conversational devices is necessary for English learners wanting to master fluent, natural-sounding speech. Additionally, it underscores the dialogic nature of language use, an element of which is what O'Keefe, McCarthy, and Carter (2007) call listenership, arguing that "to be good at

communicating and interacting, learners need to be able to show listenership and engagement just as much as they need to be able to make a point, tell a story, comment on the world around them" (p. 157).

I have come across one textbook for Russian language learners that provides a similar degree of detail about conversation strategies, *A kak ob etom skazat'?* [And How Do You Talk About That?], by Volodina (2008). The textbook intends "to acquaint [learners] with the linguistic means (language chunks and particles) used in conversational speech to relay additional subjective meaning, and to activate them in concrete communicative situations" (Volodina, 2008, p. 2). For example, the section 'Expressing agreement/disagreement with an opinion (appraisal, supposition, etc.)' begins with a series of dialogues containing targeted words and phrases (p. 141):

> – *Katia prekrasno igraet na fortep'iano.* [Katya plays the piano very well.]
>
> – ***Eshche by**! Ona uchenitsa odnogo iz izvestneishikh pianistov.* [**You bet!** She's
>
> a student of one of the most famous pianists.]
>
> – *Dumaiu, chto ei prikhoditsia mnogo rabotat'*. [I think she has to work a lot.]
>
> – ***To est'**! (= ne prosto mnogo, a ochen' mnogo, vy ochen' tochno otsenili situatsiiu)* [**Absolutely!** (= more than a lot, you gave your opinion very precisely)]
>
> – *No eto ved' neobkhodimo.* [But it's necessary, you see.]
>
> – ***Konechno**, chtoby stat' khoroshim muzykantom, odnogo talanta malo, nado ochen' mnogo i uporno zanimat'sia.* [**Of course**, in order to become a good musician, talent isn't enough, you have to put in a lot of hard work.]

Presumably this is an invented dialogue, which would account for its dryness. Adapting corpus data would result in a much more interesting and natural-sounding conversation, as in this excerpt from the Russian National Corpus:

> – *Aga, bilet 250 rublei stoil, no zato kakie tam sladen'kie mal'chiki byli!* [Yeah,
>
> the ticket cost 250 rubles, but the boys there were so cute (lit., sweet)!]
>
> – *Nu ty s kem-nibud' zazhgla, ia nadeius'…* [Well I hope you found someone to flirt with…]
>
> – *Kha!* ***Eshche by**!* [Ha! **You bet!**]

> Source: Microdialogues // From Ulianovsk University materials, 2007

Following the dialogues, *A kak ob etom skazat'?* provides a table of the target phrases with commentary on usage. *Eshche by* [you bet], for example, is used when "the speaker does not concede that the situation/person/subject could be appraised any other way." (Volodina, 2008, p. 143). The section closes with awareness-raising and discourse completion tasks, followed by several instances of the phrases in literary contexts. Use of authentic literary passages is common in RFL textbooks, given Russia's rich literary tradition, but there is a place in those materials for spoken language excerpts as well. While it contains a wealth of contextualized phrases and lexical chunks that are used widely in conversational Russian, there is no reference to how the content of *A kak ob etom skazat'?* was selected – no mention of corpus data or frequency lists of any kind. Still, this resource is a step in the right direction. Like many RFL textbooks published in Russia, it contains a wealth of useful information, but the presentation leaves much to be desired—complex grammatical terminology is used, and the layout is text-dense. Only advanced users will likely have the proficiency needed to parse this text.

Advancing in Russian through Narration by Pavlenko and Hasko (2008) is the only available resource for Russian language learners, to my knowledge, that uses a corpus-informed approach. The corpus in question consists of narratives told by native speakers of Russian and American learners of Russian with advanced proficiency. The authors analyzed the differences between the native speaker and non-native speaker narratives, then produced explanatory materials and exercises addressing those differences. *Advancing in Russian through Narration* provides an excellent example of how learner corpora can be used in the development of language materials. Nesselhauf (2004) gave several reasons for the use of learner corpora in pedagogical contexts, arguing that they "can help to decide what features should be particularly emphasized in teaching or even lead to the introduction of hitherto neglected elements (such as certain formulaic sequences, for example)" and can be "used to provide examples of typical mistakes and typical cases of overuse and underuse in teaching and in reference materials" (p. 139). Pavlenko and Hasko (2008) used a learner corpus to accomplish just those goals. The corpus revealed, among other things, that the American learner narratives lacked formulaic sequences typical of native speaker

introductions and conclusions. Additionally, the researchers found that learners of Russian tend to overuse adjectives and adverbs in describing emotions, where native speakers show a much greater preference for emotion verbs. Such research makes clear the benefits of using learner corpora to inform pedagogical materials. Regrettably, Russian language learner corpora are rare. These tools are needed and their creation should be a priority for researchers. But, as Nesselhauf (2004) noted, although learner corpora have had little impact in the field of English (as in Russian) language learning materials, "systematic learner corpus research has been carried out and the results have been used to compile or improve" learner dictionaries (p. 137). As mentioned previously, such resources have not yet been developed for Russian language learners.

The corpus-based materials examined here contain a variety of approaches to working with authentic language data, which are summarized in Table 1. As the table shows, for the most part the textbook presentation does not vary greatly from that of traditional language materials, as the corpus data are often seamlessly integrated into the content and activities.

Conclusion

In this paper, I reviewed corpus-based approaches to materials development and offered suggestions on how to use corpus data to inform RFL textbooks. In that context, corpora can assist with: selecting vocabulary based on frequency and description (in the form of lexicogrammatical profiles); identifying high- and medium-frequency chunks for presentation (including social routine formulae, conversation strategies, discourse organizers, and idioms); determining what grammatical structures are most used and in what contexts; and replacing scripted dialogues with natural spoken discourse. Like Conrad (2000), my goal is not to criticize currently available RFL materials, but rather "to emphasize that textbooks of the 21st century can be based on a more accurate analysis of language use" (p. 557). Corpus linguistics is a relatively new field, but its applications can have far-reaching consequences for language learners. Furthermore, the use of corpora can be viewed as a more efficient extension of what teachers have always done: "plucked written texts out of the contexts in which they were originally produced and imported them into the classroom, carefully selecting and mediating them for their students" (O'Keefe, McCarthy, & Carter, 2007, p. 27). The analysis of real-life language usage has been

made simpler, thanks to the continuing development of corpora, both general and specialized, and their increasingly user-friendly interfaces. Language researchers and instructors need to take these advances into account when preparing materials, syllabi, and lessons, in order to ensure their relevance and accuracy. As Reinhardt (2010) noted, teachers often rely heavily on the textbook to guide instruction, thus "corpus linguistics might be more influential in L2 instruction if it can influence the design of the materials with which teachers teach" (p. 246). Such an outcome would be beneficial to both language instructors and language learners, and should be a priority in future materials development.

Works Cited

Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Basingstoke, UK: Palgrave Macmillan.

Carter, R., Hughes, R., & McCarthy, M. (2000). *Exploring grammar in context: Upper intermediate and advanced.* Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (1997). *Exploring spoken English*. Cambridge: Cambridge University Press.

Chang, C., & Kuo, C. (2011). A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes, 30*(3), 222-234.

Conrad, S. (2000). Will corpus revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, *34*, 548-560.

Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 67-85). Amsterdam: John Benjamins Publishing.

Conrad, S., & Biber, D. (2009). *Real grammar: A corpus-based approach to English*. White Plains, NY: Pearson Education.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*, 393-417.

Gavioli, L., & Aston, G. (2001). Enriching reality: language corpora in language pedagogy. *ELT Journal, 55*(3), 238-246.

Halliday, M. A. K., Teubert, W., Yallop, C., & Čermáková, A. (2004). *Lexicology and corpus linguistics: An introduction.* London: Continuum.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4*, 1-16.

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring*. Boston, MA: Heinle.

Lazar, G. (2003). *Meanings and metaphors: Activities to practise figurative language*. Cambridge: Cambridge University Press.

Mauranen, A. (2004). Spoken corpus for an ordinary learner. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 89-105). Amsterdam: John Benjamins Publishing.

McCarthy, M. (2004). *Touchstone: From corpus to course book*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone: Student's book 2*. Cambridge: Cambridge University Press.

McCarthy, M., O'Dell, F., & Shaw, E. (1997). *Vocabulary in use upper intermediate*. Cambridge: Cambridge University Press.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Amsterdam: John Benjamins Publishing.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Pavlenko, A., & Hasko, V. D. (2008). *Advancing in Russian through Narration*. University Park, PA: CALPER Publications.

Reinhardt, J. (2010). Viewpoints: The potential of corpus-informed pedagogy. *Studies in Hispanic and Lusophone Linguistics, 3*, 239-252.

Rifkin, B. (1996). *Grammatika v kontekste: Russian grammar in literary contexts*. Boston: McGraw-Hill.

Simpson-Vlach, R. C., & Leicher, S. (2006). *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. Ann Arbor: University of Michigan Press.

Tomlinson, B. (1998). Introduction: Principles and procedures of materials development. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 1-30). Cambridge: Cambridge University Press.

Varley, S. (2009). I'll just look that up in the concordancer: Integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning, 22*(2), 133-152.

Volodina, G. I. (2008). *A kak ob etom skazat'?: Spetsificheskie oboroty razgovornoi rechi*. Moscow: Russkii Iazik.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, *21*, 3-25.

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