Chapter 10
Innovative Methods for LSP-Teaching: How We Use Corpora to Teach Business Russian

[AU1] **James Wilson, Serge Sharoff, Paul Stephenson, and Anthony Hartley**

10.1 Introduction

In the last two decades much work has been carried out on corpus development; many large corpora, both general reference and specialised, have been created and user-friendly tools that allow non-specialist users to perform complex corpus queries and even build their own corpora have been developed. Consequently, we have seen an influx of corpus-based resources, in particular dictionaries, such as the *Collins Cobuild English Language Dictionary* (1995) and the *Cambridge Advanced Learners Dictionary* (2008), and grammars (Biber et al. 1999; Willis 2004), and the *indirect* impact of corpora in language learning and teaching has been considerable. Conversely, the *direct* impact of corpora in the language learning classroom has been less impressive, especially in Foreign Language Teaching (FLT), and we might argue that little progress has been made since Tim Johns' work on Data Driven Learning (DDL) in the 1980s and 1990s (see, for example, 1991a, b, 1993). That is, few tutors make active use of corpus-based exercises or use corpora in their classes. Despite the abundant general literature on corpora (e.g. McEnery and Wilson 1996; Biber et al. 1998; Adolphs 2006; Anderson and Corbett 2009; O'Keeffe and McCarthy 2010) and the rapidly growing literature on the application of corpora in

J. Wilson (✉)
Department of Russian and Slavonic studies, Centre for Translation Studies,
University of Leeds, Leeds, UK
e-mail: j.a.wilson@leeds.ac.uk

S. Sharoff
Department of Modern Languages and Cultures, Centre for Translation Studies,
University of Leeds, Leeds, UK
e-mail: s.sharoff@leeds.ac.uk

P. Stephenson • A. Hartley
Centre for Translation Studies, University of Leeds, Leeds, UK
e-mail: a.hartley@leeds.ac.uk

E. Bárcena et al. (eds.), *Languages for Specific Purposes in the Digital Era*,
Educational Linguistics 19, DOI 10.1007/978-3-319-02222-2_10,

language teaching (e.g. Sinclair 2004; Aston et al. 2004; Hidalgo et al. 2007; Aijmer 2009; Bennet 2010; Reppen 2010; Harris and Moreno 2010; Frankenberg-Garcia et al. 2011), the use of corpora for pedagogical purposes has not become an established practice. In fact, many language tutors are not sure what corpora are (Mukherjee 2004) or what they may be used for (McCarthy 2008). Tutors who [AU2] use corpora in their teaching form a small and generally expert community, and even they tend to use corpora predominantly as a source of reference for checking their language intuition and therefore do not exploit the available technology to the full. Most work has been done on the use of corpora in English Language Teaching (ELT) (e.g. Braun et al. 2006; O'Keeffe et al. 2007; Bennett 2010; Campoy et al. 2010; Frankenberg-Garcia et al. 2011) and, increasingly, on learner corpora (Aijmer 2009), while FLT has remained almost untouched by the corpus "revolution". Furthermore, even for ELT many works in the literature reflect corpus linguists' research interests in language use and many corpus-based examples are chosen because they are linguistically interesting, not because they are pedagogically relevant.

This is unfortunate in that a corpus-based approach to language learning and teaching has many benefits for both tutors and their students. Corpora can have a considerable positive impact on Language for Specific Purposes (hereafter, LSP) teaching, a rapidly growing market that is particularly relevant to the linguistic needs of industry but for which there is a lack of conventional printed teaching materials. Using corpora alongside established teaching practices, we can offer tuition in several LSP domains and tailor our curricula to the needs of industry and increase the employability of our graduates. Our case study is a business Russian class offered at the University of Leeds; we demonstrate how we use corpora in our language classes and in what ways a corpus-based approach is particularly well suited to business Russian and other LSP subjects.

The business Russian class is part of an applied Russian language module that is taken by students in their third year of study (having returned from a term or year in Russia) and it prepares students for work in industry. Several topics are covered on the course; these include: formal letter writing, CVs and covering letters, business presentations, sales pitches, and official telephone conversations. While such a module is of particular relevance to students intending to work in Russia upon completing their degree, tutors found it hard to structure and deliver because there are few relevant teaching materials. Similarly, students expressed that, unlike for their other subjects, there was not a (single) course book containing all or most of the materials relevant to the business Russian module. Students need access to notes on business-specific terminology and conventions in formal etiquette and style as well as sample official documents of various kinds (CVs, covering letters, letters of complaint, etc.) – documents which students are expected to produce in class tests and for homework assignments (as well as in "real-world" scenarios beyond the degree). Before introducing corpus-based materials to the module we relied heavily on web-based material, which was gleaned from many Russian-language websites and needed to be adapted for pedagogical purposes, and we also studied materials

from books on business Russian and official written discourse produced in the Czech Republic (and intended primarily for Czech learners of Russian).[1] Although concerns over the lack of adequate teaching materials were voiced specifically about business Russian, we might argue that there is a dearth of printed materials for other LSP subjects and for other languages, especially though not exclusively for less commonly taught ones.

Two principal skills that we look to train on the business Russian module are: (1) vocabulary acquisition and (2) register recognition and differentiation. Although students taking the applied Russian module have spent 9 months in Russia, the business Russian class is their first real experience of formal writing. The core lexicon of business Russian is different to that of other genres that students are already familiar with, and many important discipline-specific words and phrases are not found in standard dictionaries or other language learning resources that students use. Furthermore, even at the (upper-)intermediate level, students are not fully competent in recognising and successfully differentiating between different varieties of written Russian. They tend to struggle in formulating official documents, as they are unfamiliar with formal salutation and valediction, higher-level grammatical and syntactic constructions and generic or discipline-specific formal clichés. As the students already possess generic language learning skills and have a solid grounding in general written Russian (grammar, vocabulary and syntax) when they enrol on the course, the task of acquiring a discipline-specific lexicon and specific morphosyntactic constructions should be straightforward, provided that there is a reliable means of identifying the vocabulary and register of a particular LSP domain. In fact, assessment shows that students encounter few problems in learning terminology specific to business Russian; however, tutors' task of compiling lists of discipline-specific terminology and constructions, given the lack of resources, is time-consuming and laborious.

Our aim is therefore to use corpora to facilitate vocabulary acquisition and to enhance students' ability to recognise formal expressions and use them in an appropriate context. We describe herein how corpora can be used to compensate for the lack of printed materials, how they facilitate tutors' task of designing relevant materials and how a corpus-based approach is particularly relevant to LSP teaching. This study contributes to the literature on the application of corpora in language learning and teaching and, more specifically in FLT and in LSP domains, by showcasing the work we have done, both from a technological perspective to enhance possibilities for manipulating corpora for pedagogical purposes and from a pedagogical perspective to develop materials to utilise the technology more fully. We demonstrate how we have simplified our existing corpus-based tools and enhanced them to facilitate and enhance teaching and research across the arts and

[1] These books include several recent publications that focus on Russian in the spheres of commerce, law, international relations and administration (Dynda and Dyndová 1998; Kupcevičová and Vilímek 2006; Mrověcová 2007; Dlouhá 1998; Golčáková 2008; Rezková 2008).

humanities. More specifically, we focus on the application of our tools in the business Russian class and describe how tutors and students may use our tools to:

1. generate frequency lists of business-related keywords;
2. automatically extract lists of business-related collocations;
3. search for affixed and compound forms;
4. automatically classify texts according to genre and domain;
5. rank concordance lines according to their level of difficulty;
6. collect their own corpora of business texts and official documents and analyze them using the functions that are available in our interface.

We then present examples of how the individual search options in the IntelliText interface (see Sect. 10.2) are used to train vocabulary acquisition and register recognition and differentiation – skills that are essential for any LSP domain – and to enhance students' language competence more generally. Finally, while our focus is on technological advances in corpus development and the application of corpus-based tools in LSP teaching, part of our chapter is devoted to material design and reports on our efforts to create corpus-based language exercises to complement and support our tools and that allow students to exploit the available technology more fully.

10.2 Simplifying and Enhancing Corpora for Teaching

In most cases, corpora are not conceived or designed for the language learning classroom but are usually intended for a specialist, technologically savvy target audience. This means that for pedagogical purposes corpora must be simplified and adapted in various ways in order to become established in the mainstream and to be accepted by language tutors. Large collections of electronically-stored texts, especially those that are not annotated, are of little practical value to language tutors or their students. The same holds for the output of a corpus search: other than for reference, a list of concordance lines is of little use. A corpus that is parsed and tagged is more useful; language learners can make more elaborate corpus queries and search for, say, a noun in a particular case, a verb in a particular tense or a multi-word expression. However, to perform more complex searches language learners often need to be familiar with complex regular expressions and competent in using computer syntax, as many interfaces require grammatical information to be typed in as a string code. Assume that a language learner wants to see which adjectives are commonly used in the expression *to make an impression* between the words *an* and *impression*. In our former interface (http://corpus.leeds.ac.uk/internet.html) he or she would have needed to enter the following string code (CQP syntax):

```
[lemma = "make"] [pos = "DT"] [pos = "JJ"] [word = "impression"]
```

The string tells the regular expression processor to look for any form of *make*, a determiner (DT) – this allows for an article before the adjective; a more specific search for examples that occur with the indefinite article could be achieved by replacing [pos="DT"] with [word="a"] – an adjective (JJ) and the word *impression*. Even one-word searches in which the search word is in a particular grammatical case may require a complex code in highly inflected languages. Take the Russian word *работа* "work" in the dative singular case (*работе*), for example. Users would need to know the following string code in order to search for this form:[2]

[lemma="работа"&pos="Ncfsgn*"] (or [lemma="работа"&pos="N..sg.*"])

Without clear and comprehensive instructions, these searches would be beyond the competence level of most untrained users and such a counter-intuitive method overwhelms many non-specialists. Moreover, until recent years the main target audience for the outputs of research on the development and use of corpora has been the immediate specialist community of computational and corpus linguists; therefore, user documentation is in some cases inaccessible to users from other disciplines.

A second and equally serious problem concerning the use of corpora for pedagogical purposes is that the content of many of the examples that the concordancer generates is too difficult for all but advanced students. This is certainly true in the case of languages that are taken up from the ab-initio level at British universities such as Russian, Arabic and Chinese. Or many examples that are generated are not pedagogically relevant, and in preparing their notes or exercises language tutors might have to sift through hundreds of examples to find just one or two appropriate ones. Therefore, tutors may find it quicker and easier to produce materials on the basis of their intuitions rather than taking examples from corpora. For teaching and learning purposes, tutors and their students need to be able to refine their search by filtering concordance lines according to several parameters (difficulty, genre, domain, grammatical features, and other categories). Language tutors and learners also require additional functions such as frequency lists, lists of useful collocations and the option to search for affixed forms, display their search word in various positions within a sentence, compare words and phrases within and across corpora, and so on.

It is evident from the above that there are two issues that need to be taken into account if corpora are to be enter the mainstream of language learning and teaching. First, corpus tools need to be made simple so that non-specialists can use them. Second, corpus-based tools and functions need to be developed specifically for a target audience of language tutors and learners. Language learners require specific

[2] There is a reasonably high degree of syncretism in the Russian case system and *работе* is both the dative and preposition singular form. Users could not therefore simply enter *работе* as a word form, unless they did not need to restrict their search to a particular case.

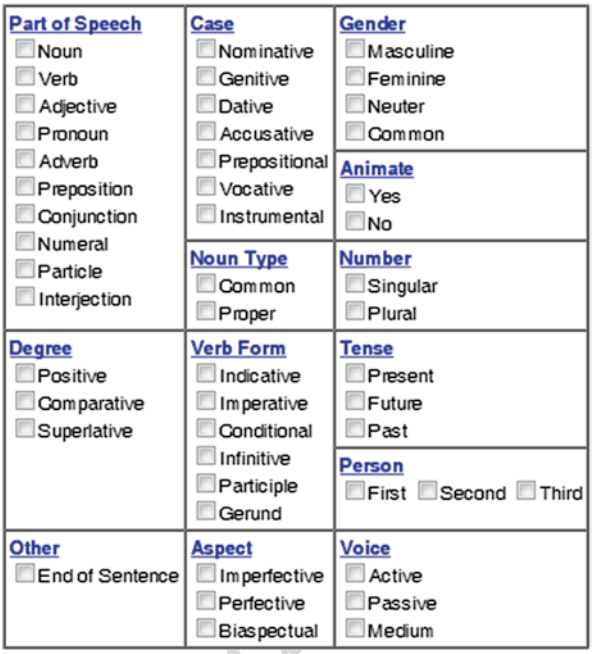


Fig. 10.1 Russian Part of Speech (POS) editor

options for building corpora and generating frequency lists, and they require a more elaborate, yet user-friendly, annotation system for classifying texts and individual concordances and collocations according to several parameters.

We have already taken steps to address these demands on the Intelligent Tools for Creating and Analysing Electronic Text Corpora for Humanities Research (IntelliText) project (Wilson et al. 2011) carried out at the Centre for Translation Studies at the University of Leeds. The broad aims of the project were (1) to make our corpora and corpus-annotating software simple to use and accessible to a wide and diverse group of users and (2) to create a versatile and intuitive interface with tools and functions aimed at enhancing and facilitating the work of teachers and researches in various areas of the humanities. Our four targeted disciplines were language learning and teaching, translation studies, linguistics and history, and we liaised with academics from these disciplines to receive their advice and guidance on the features that should be implemented in our interface. Most of our collaborators, mainly academics with little or no experience in using corpora in their teaching or research, emphasised the importance of a simple and intuitive interface that is supported by clear and comprehensive documentation including discipline-specific tutorials.

With the aim of making our interface more intuitive and user-friendly, we have implemented a part-of-speech (POS) editor that allows users to add grammatical information to their search by ticking check-boxes (see Fig. 10.1), rather than by

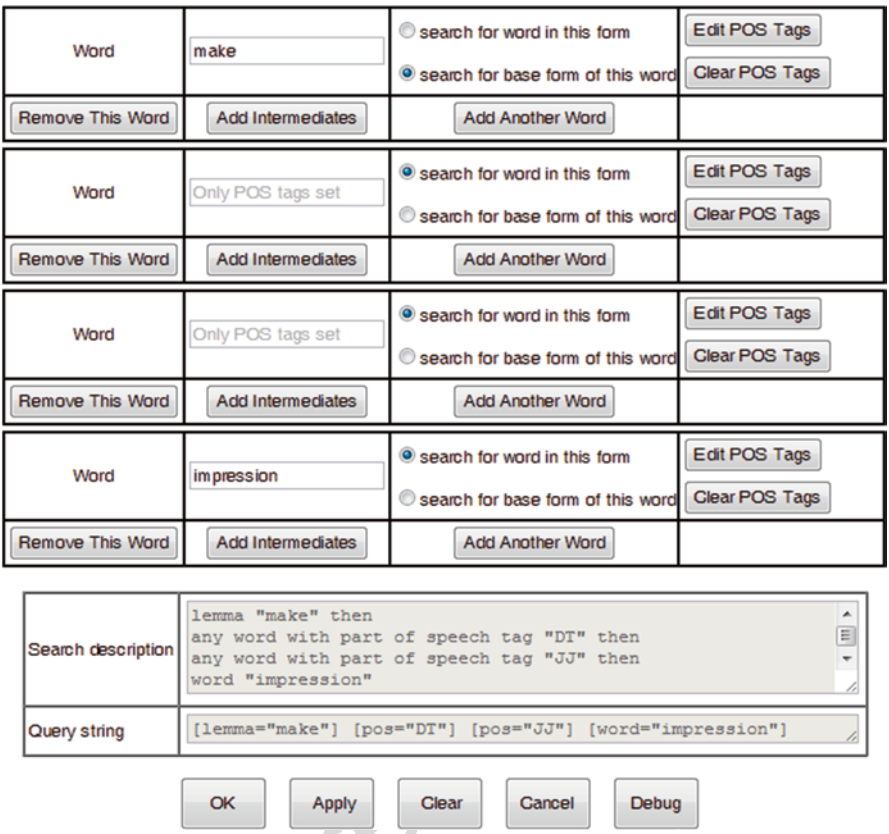


Fig. 10.2 Search builder

manually typing out complex string codes. We have also simplified multi-word searching by developing a search builder (see Fig. 10.2) that allows users to formulate complex queries, including searches involving specific and non-specific intermediates, and to specify whether their search word is a lemma or not. Other functions developed and implemented on the IntelliText project are discussed below in the context of their application in LSP teaching.

We are also tackling the second problem outlined above and, acting on the advice of language tutors, we have made several modifications to our interface that directly benefit language learning and teaching. Users may search for words in various places in a sentence, display their search words at the start, end or in the middle of a concordance line, search for affixed forms of words without needing to know string codes or symbols (see Sect. 10.3.1.3) and compare the use of two or more competing words or phrases. To support the general IntelliText user guide we have also created language-specific tutorials for English as a Foreign language, Japanese and Russian, which provide sample searches for and exercises on

linguistic issues relevant to these languages.[3] With respect to developing a more sophisticated annotation system, we are researching possibilities for classifying texts and individual concordance lines (see Sect. 4.2) according to their level of difficulty and we have compiled corpora stratified according to topics studied by students at UK institutions at levels 2 and 3 for German, Japanese and Russian. [AU4]

10.3 Corpus-Based Approaches to Language Learning and Teaching

As corpora are a fairly new addition to the teaching toolkit, many users are unaware how to use them, what to use them for or what benefits they bring to learning and teaching (McCarthy 2008). This is not surprising in that, despite recent progress in EFL, little has been done to promote corpus-based learning among the wider language learning community. This is unfortunate, as we believe that a corpus-based approach to language learning can supplement and extend the effectiveness of existing teaching materials. Using corpora, students and tutors can:

- view grammar in context;
- display complex grammatical forms not shown in conventional bilingual dictionaries;
- access hundreds of authentic examples at the touch of a button;
- view vocabulary in a broader context, extracting common and useful collocations that are beyond the remit of printed resources;
- grasp subtle differences between words and phrases;
- verify their linguistic intuition;
- achieve a better grasp of style;
- augment their vocabulary, in particular on themed topics and in specific domains;
- test controversial points of grammar and compare prescribed grammar with actual language use.

More specifically, in LSP teaching corpora can be used to train specific language skills and in some cases can go beyond the remit of printed resources. Here we look specifically at their application in the context of the business Russian module and how they may be used to train vocabulary acquisition and register recognition and differentiation. Even in the largest dictionaries and the most comprehensive grammar manuals only a few phrases, collocations and constructions are shown (in some cases out of context), and the information given may be insufficient for language learners to fully grasp how a particular word or grammatical feature is used in context. Corpora, on the other hand, are made up of hundreds of authentic language examples and collocations, thus language learners may view problematic words or grammatical forms in hundreds of sentences and in different types of text and context.

[3] http://smlc09.leeds.ac.uk/itweb/cmsimple/?Sample_Corpus_Searches_%28Russian%29

Moreover, printed materials tend to cover only the main LSP domains, as it is financially impracticable to publish books on topics that do not have a wide target audience, and they cannot represent current affairs. From corpora, however, tutors may develop materials and create ad-hoc exercises for highly specific topics or to create vocabulary lists on current topics that students can use to describe events that are taking place at the present time. In fact, tutors may create materials for any topic, provided that enough written material is available for that topic. For example, for our final-year core Russian language module we created corpora of texts on the Tsunami that occurred in Japan in March 2011 and on the Libya Crisis and the death of Muammar Gaddafi from Russian-language newspaper articles available online. We use these and other corpora, built in the same way, and exercises based on their content to augment students' vocabulary and to train students to discuss topics, both orally and in writing, using sophisticated and stylistically appropriate words, phrases and constructions. In fact, tutors may create materials for any topic, provided that enough written material is available for that topic.

10.3.1 Vocabulary Acquisition

In this section, we shall focus on three corpus-based methods of vocabulary acquisition: (1) frequency lists, (2) collocation extraction and (3) compound and affix searching. We shall also describe how tutors and students can use our tools to create DIY corpora and use these corpora to extend their vocabulary on themed topics.

The acquisition of a core, discipline-specific vocabulary can provide a solid basis for creating materials for modules such as business Russian, French for lawyers, Spanish for medics, and other LSP-based subjects, whose core lexicon differs markedly from that required for "general" language learning and many key phrases of which are not found in standard bilingual dictionaries. Tutors can quickly build their own corpora for various LSP domains, no matter how specialised, and then generate frequency lists of the most common words and phrases in these corpora. The benefits of this approach were clearly demonstrated by Butler (1974). Butler used frequency lists extracted from a small corpus (94,000 words) of systematically selected chemistry papers to "permit second-year chemistry undergraduates (or postgraduates), with no previous knowledge of German, to read papers from German chemical journals for comprehension and, where necessary, for translation" (50). Using exercises built around the most frequent words and collocations from his "chemistry" corpus, Butler aimed to equip his students with the requisite reading skills within 10 teaching hours (plus 20 h of independent study). The results were "gratifyingly successful" (53). Butler's methodology can be extended to other languages and to other LSP domains.

We are currently testing a similar methodology on the AHRC-funded ReadingCorp (A corpus-based approach to achieving reading competence in Russian) project that is being carried out at the Universities of Leeds and Sheffield. Building on previous research carried out on the CEELBAS-funded Russian for

Research project,[4] on which we indentified that existing postgraduate (PG) language provision at many institutions is inadequate and many potentially good researchers are consequently slipping through the net, we aim to test whether a corpus-based and vocabulary-oriented approach can be used to provide PhD students with little or no knowledge of Russian the necessary reading skills to comprehend primary texts in their area of research. Keyword lists are extracted from two specialised corpora: a general academic corpus made up of academic articles in areas of interest specified on the Russian for Research project and a smaller corpus of texts directly relevant to the students' research topic. Besides producing an extensible methodology and discipline-specific exercises, we shall also deliver an online grammar of Russian specifically for researchers in which emphasis lies firmly on the identification of common grammatical structures in academic texts selected from corpus-based frequency data.

Quick acquisition of discipline-specific vocabulary is highly desirable in LSP teaching, as in the above-mentioned case of PhD students who need to learn a foreign language from scratch for their research. Intensive undergraduate ab-initio programmes are not suited to researchers' language needs; much of the material covered at Level 1 is of little relevance to researchers, while many important points of grammar and language use (from the researcher's perspective) are not covered. Gaining the necessary reading competence to understand academic texts is quite straightforward: academic texts are characterised by lexical and stylistic repetition; therefore, corpus-derived frequency lists can be particularly effective, as Butler demonstrated. The same holds for grammar: teaching can be structured around the grammatical constructions that occur most frequently in academic texts and training need not involve production and correct use of these forms, but simply the ability to recognise them and know what purpose they serve. As specialised intensive language for research courses are not practicable at many institutions (especially in the case of less commonly taught languages) and there are concerns over the sustainability of established courses, new, cost-effective modes of delivery are highly sought after. A corpus-based, vocabulary-oriented approach is a realistic and inexpensive way of training or improving reading skills quickly and effectively.

10.3.1.1 Frequency Lists

Until recently, tools for extracting frequency and keyword lists from corpora belonged to the exclusive domain of computational linguists and non-specialists needed technical assistance to produce them. Nowadays, however, a standard feature of many corpus interfaces is a built-in and easy-to-use frequency list generator. Users can use our tools to generate lists of single- or multi-word terms and they can generate both frequency and keyword lists from the corpora available through our

[4] http://www.ceelbas.ac.uk/research/networkprojects/funded2007 (project CN07SF-1).

t1.1 **Table 10.1** Russian keywords sorted by the Log-likelihood score of their significance

t1.2	Single words	Multiwords
t1.3	банк (bank)	ценные бумаги (securities)
t1.4	предприятие (enterprise)	юридическое лицо (organisation)
t1.5	кредит (credit)	денежные средства (monetary assets)
t1.6	договор (contract)	федеральный закон (federal law)
t1.7	товар (product)	заработный плата (salary)
t1.8	рынок (market)	бухгалтерский учет (accounting)
t1.9	финансовый (financial)	земельный участок (land plot)
t1.10	налог (tax)	акционерное общество (public company)
t1.11	страхование (insurance)	рынок ценных бумаг (capital market)
t1.12	цена (price)	основные средства (fixed-capital assets)
t1.13	учет (accounting)	фондовая биржа (stock exchange)
t1.14	ценный (valuable)	процентная ставка (interest rate)
t1.15	денежный (money-adj)	фондовый рынок (stock exchange)
t1.16	имущество (property)	арбитражный суд (arbitrage)
t1.17	налоговый (tax-adj)	недвижимое имущество (real estate)
t1.18	прибыль (profit)	система управления (system of management)
t1.19	государственный (state)	налоговые органы (tax authorities)
t1.20	страховой (insurance-adj)	договор страхования (insurance contract)
t1.21	стоимость (cost)	инвестиционный фонд (investment fund)

interface or from corpora that they upload to it. A frequency list is understood here as a list of the most common words or lemmas in a corpus; a keyword list as a list of words that appear in a corpus more often than we would expect by chance and that are extracted by statistical tests that compare their frequency against expected frequencies derived from larger reference corpora (as, for example, the words and phrases in Table 10.1). While the output of keyword lists is useful only for reference and passive language acquisition, tutors can create exercises around the output in order to promote active language acquisition (see Sect. 10.4).

Single-word terms from specialised corpora are detected by Log-likelihood scores for their frequencies against a reference corpus (Rayson and Garside 2000). An adaptation of the commonly-used Bootcat algorithm (Baroni and Bernardini 2004) is used for the extraction of multi-word terms. For example, we can start with the frequencies of Russian words in the overall Internet corpus (Sharoff 2006) and the frequencies from the business corpus to get the list of keywords. Then the list can be extended by finding commonly used sequences of several words that contain at least one of the words in the keyword list. The importance of multiword units can be highlighted by non-compositional expressions, like *ценные бумаги* (securities) which literally means "valuable papers". Users will be eventually be able to specify the length of the phrases generated in the lists and select to view keyword lists of specific parts of speech such as verbs, adjectives and nouns.

10.3.1.2 Automatic Extraction of Collocations

Tools for indentifying and displaying collocations in corpora are now well established in many interfaces. There is an unresolved issue with regard to the most appropriate score by which to rank the collocates (Evert and Kren 2001), but the Log-likelihood score (Dunning 1993) is generally considered as a reasonable choice, as it takes into account the ratio of frequencies as well as the actual number of examples, Our tools can rank collocations either by their joint frequency or by three statistical scores: Log likelihood, Mutual Information and T-Score. All three statistical methods are used to determine the statistical significance of two words occurring together; the default sorting order in the collocation tables is by log likelihood.

Language learners often produce odd collocations, usually because of interference from their native language (L1), and collocational information in most dictionaries is scant – obviously because space is limited (even in larger dictionaries). The automatic extraction of collocations is therefore another corpus-based function that can be used to good effect in learning and teaching and it exceeds the capabilities of existing printed resources. In keeping with the business theme of this paper, let us look at two examples from Russian: adjective + *рынок* "market" and *произвести* "to make, produce" + noun. Our tools allow users both to restrict the grammatical tag of the collocates and to select their position in relation to the search word; therefore, we can tell the corpus to show adjectives on the left of *рынок* (see Fig. 10.3) and nouns on the right of *произвести* (see Fig. 10.4). We used the Russian Business Corpus available via our interface for both queries. To get a better idea of how these collocations are used in context students can then click either on the collocations in the "Pair" column or on the number beside them in the "Count" column of the table and view a list of concordances.[5] A more detailed account of how we use the collocation search function on the business Russian module is given in Sect. 10.4.

10.3.1.3 Compound and Affix Searching

As the aims of IntelliText were to create a user-friendly and simple search interface, we have implemented a search type specifically intended for affixes that allows users to either enter an affix and see in which words that affix occurs or enter a lemma to generate affixed forms of that lemma (see Fig. 10.5). Our Affix Search function is a particularly valuable resource for vocabulary building in many languages, particularly those for which affixation is a productive source of word formation, such as for the Slavonic languages in the case of verbal prefixation. The lists in Fig. 10.6 show the outputs of both type of affix search.

[5] Users generate ten concordances by clicking on the collocation and a list of all occurrences of the collocation by clicking on the number.

Pair	Count	F1	F2	LL	MI	T
фондовый рынок	302	3543	5310	697.5	8	17.31
российский рынок	185	17829	5310	230.38	4.96	13.16
вторичный рынок	94	838	5310	229.7	8.39	9.67
мировой рынок	125	5542	5310	203.17	6.08	11.01
валютный рынок	110	4543	5310	182.61	6.18	10.34
внутренний рынок	111	4954	5310	179.98	6.07	10.38
внебиржевой рынок	47	266	5310	126.24	9.05	6.84
первичный рынок	58	1506	5310	109.67	6.85	7.55
финансовый рынок	85	14627	5310	81.79	4.12	8.69
страховой рынок	58	7026	5310	65.47	4.63	7.31
организованный рынок	27	356	5310	60.37	7.83	5.17
международный рынок	56	9956	5310	52.95	4.08	7.04
внешний рынок	39	3890	5310	47.66	4.91	6.04
межбанковский рынок	22	376	5310	46.24	7.45	4.66
междилерский рынок	12	19	5310	41.79	10.89	3.46

Fig. 10.3 Top ten adjective collocates to the left of *рынок* in the Russian Business Corpus

Pair	Count	F1	F2	LL	MI	T
произвести впечатление	332	6147	10171	908.81	9.28	18.19
произвести фурор	39	6147	115	155.34	12.65	6.24
произвести обыск	53	6147	1588	145.07	9.31	7.27
произвести эффект	66	6147	5943	144.1	7.72	8.09
произвести продукция	71	6147	11862	133.14	6.83	8.35
произвести посадка	30	6147	2839	64.71	7.65	5.45
произвести замена	30	6147	3200	62.92	7.48	5.45
произвести расчет	38	6147	9881	62.88	6.19	6.08
произвести сенсация	20	6147	938	50.18	8.66	4.46
произвести выстрел	23	6147	3694	43.54	6.89	4.76
произвести платеж	19	6147	2764	36.9	7.03	4.33
произвести расход	22	6147	8184	32.51	5.67	4.6
произвести ремонт	20	6147	5750	32.09	6.05	4.4
произвести пуск	14	6147	1012	32.08	8.04	3.73
произвести оплата	20	6147	5958	31.74	5.99	4.4

Fig. 10.4 Top ten noun collocates (Collocates are displayed in their lemma form) to the right of *произвести* in the Russian Business Corpus

this figure will be printed in b/w

Fig. 10.5 The Affix Search page

this figure will be printed in b/w

Lemma	Count	Forms	Lemma	Count	Forms
передать	10248	65	считать	41768	78
перестать	9581	47	прочитать	6348	59
перейти	7949	37	предпочитать	4169	53
перевести	5892	57	рассчитать	3020	48
переходить	5856	63	почитать	2365	65
передавать	5424	60	посчитать	1666	40
переживать	4589	63	подсчитать	875	40
пережить	4379	63	пересчитать	609	34
перенести	3234	51	зачитать	497	37
переводить	2944	63	перечитать	494	27
переставать	2612	27	вычитать	478	46
перебить	2580	43	сосчитать	461	29
переехать	2351	38	просчитать	425	37
переносить	1999	51	причитать	352	22
перечислить	1873	43	дочитать	334	26

Fig. 10.6 Sample outputs of an Affix Search: verbs beginning with *пере–* (on *left*) and prefixed forms of *читать* (on *right*)

Unfortunately, the searching mechanism cannot distinguish between genuine affixed forms and non-affixed forms containing the same combinations of letters. For example, a search for the prefix *over* looks for all words beginning with over and so finds words such as *overt* and *overly* as well as genuine affixed forms such as *overdone* and *overlay*. Similarly, a search for affixed forms of the verb *do* generates words like *Nintendo* and *torpedo*. Users can try to improve the accuracy of the output by selecting the "Highlight likely non-affixed forms" option, which highlights in red forms that are not lemmas when the affix is removed. For example, *Nintendo* and *torpedo* can be filtered out, since the forms *Ninten* and *torpe* are not lemmas.

However, this function is unreliable for many languages and retagging is needed to knock out non-affixed forms from affix searches.[6]

Though called an Affix Search, this function can also be used to display lists of compound forms such as ***работо****датель* "employer", ***работо****дательница* "female employer", ***работо****способность* "work capacity"; *клиенто****ориентированность*** "client-centeredness", *рентo****ориентированность*** "rent-centeredness", *бизнес****ориентированность*** "business-centeredness", *англо****ориентированность*** "English-centeredness".

10.3.1.4 Using DIY Corpora for LSP Teaching

Besides using existing corpora to generate frequency lists, lists of collocations, affixes and compounds, users may also compile their own collections of texts which are automatically lemmatised and grammatically tagged by our tools. DIY corpora are useful in both language and translation for specific purposes, in cases when tutors, students or translators need to work with specific types of language. In combination with the Frequency List Generator, DIY corpora offer a quick and easy means of extracting a core vocabulary list for any LSP domain. Users simply need to upload a set of texts, either as plain text, Microsoft Word or comprised files and they can then make full use of all the features built into our interface. With regard to the business Russian module, students may collect advertisements, CVs and covering letters, letters of complaint, political leaflets, etc., and then extract keywords from their corpus or look up specific words in concordance and collocation searches. This is a major benefit of corpus-based learning: while the lexicon of such a breadth of highly-specific domains cannot be represented in printed language-learning resources, which often take years to produce and generally cater for wider groups of users, corpus-derived frequency lists can be created within hours and, more importantly, can cater for the individual. Corpora also give students greater autonomy and allow them to extract their own frequency lists from and perform other searches on corpora that they have compiled – an activity that can be both fun and fulfilling.

10.3.2 Register Recognition and Differentiation: "Tagging for Teaching"

Many modern corpora are enriched with grammatical and domain information that allows users to refine their searches and look for specific word forms, select texts according to parameters like genre and topic and classify them in several other ways. Besides morphological and semantic tagging, now customary in many

[6] In the case of Russian, words like *президент* "president" and *предприятие* "enterprise" are correctly highlighted as non-affixed forms, as *зидент* and *дприятие* are not lemmas. On the other hand, words like *сон* "sleep; dream" and *сад* "garden" are not highlighted as non-affixed forms, as *он* "he" and *ад* "hell" are lemmas.

modern corpus interfaces, several other types of tagging can be used to enhance language learning. In this section, we look at three types of annotation that are relevant to LSP-teaching: (1) genre classification, (2) difficulty ranking and (3) the classification of texts according to the grammatical forms that they contain.

10.3.2.1 Genre Classification

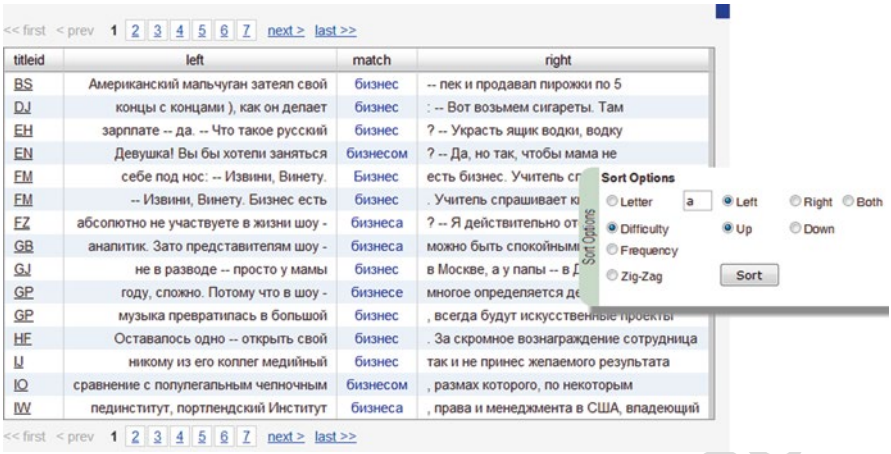
Genre classification, a system by which genre categories can be automatically assigned to texts in non-annotated corpora, is a major advance in corpus development and allows us to show stylistic variation across texts. On the basis of previous research in which parameters for automatic genre detection and classification of texts from the Web (Sharoff 2010; Sharoff et al. 2010), we have been able to classify texts into such general genre categories as news items ('reporting'), legal texts ('regulations'), FAQs, advice ('instruction'), promotional materials ('advertising'), listings ('information') and everything else ('discussion').

Modern text classification (which covers genre and topic classification) is based on (1) detecting features which represent each individual document and (2) applying Machine Learning algorithms using a collection of documents with known labels. We used the Weka toolkit for the latter task (Witten and Frank 2005). Given a topically diverse document collection, the genre of documents can be successfully encoded by character n-grams, i.e., sequences of n characters (Mason et al. 2010; Sharoff et al. 2010). For example, the genre of news reports in English can be described by such n-grams as: rday week sday annou, which generalise over references to dates *(yesterday, Tuesday, Wednesday, Saturday),* while the genre of research articles can be described by such n-grams as s by; ly l; eris; naly, which generalise over some of the features of the formal language (including the greater use of adverbs and the passive voice). Such features are not absolute; Machine Learning algorithms (such as Naïve Bayes or Support Vector Machines) learn the strength of association between individual features and the labels, reaching the accuracy of 70–80 % on a topically diverse corpus.

In a similar fashion, for business Russian, we have been able to apply formality tags so that business clichés can be extracted to help students differentiate between formal and neutral language use as well as to help them achieve a better understanding of register and to allow them to compile their own lists of generic phrases used in official writing. Such stylistic tagging has a much wider application and can be used to display other features such as non-standard or regional use.

10.3.2.2 Difficulty Ranking

As we mentioned earlier (see Sect. 10.2), many language tutors have highlighted that the use of corpora in their language classes is limited in that students find it hard to understand the content of the concordance lines. This is not surprising: most corpora are made up of authentic texts and were not designed with the language



this figure will be printed in b/w

Fig. 10.7 Difficulty ranking in the IntelliText interface

learner and difficulty levels in mind. Most corpora were created for linguistic purposes, and while it is possible to develop tools to exploit these corpora for specific audiences, it is not possible to change their contents, and it is obviously time-consuming to create new pedagogically oriented corpora made up of graded texts and stratified according to content and other phenomena relevant to language learning. As a consequence, in the case of many languages, only advanced language learners have full access to corpus-based language learning, and even then tutors often need to sift through many examples to find just one or two appropriate ones.

Some recent advances have been made in this respect. Sharoff et al. (2008) established parameters for assessing the difficulty of texts and individual sentences in Russian (and other languages) and mapped the result to the CEFR levels.[7] Methods have also been developed for automatically ranking concordance lines according to their difficulty by measuring lexical and syntactic complexity and sentence similarity (Segler 2007; Kilgarriff et al. 2008; Kotani et al. 2008). As additional options for sorting the concordance lines we have integrated a relatively straightforward lexical match using the frequency-based CEFR-annotated lists (Kilgarriff 2010). This means that the difficulty level of the concordance lines displayed can be matched to the experience level of the students. Our user interface provides simple intuitive components, such as sliders, to ensure that setting the desired level of difficulty is easy. Even though, like other automatic processes, difficulty tagging is not completely accurate it still makes the tutors' task of selecting appropriate examples from the concordance lines much easier. Like all the processing features in our system difficulty tagging has been implemented in a plug-in fashion so that it can easily be replaced as better methods are developed (Fig. 10.7). [AU5]

[7] http://www.coe.int/t/dg4/linguistic/cadre_en.asp

10.3.2.3 Text Mining for Specific Grammatical Forms

Corpus-held texts may also be categorised according to grammatical forms and users may select to view texts that contain specific grammatical features. Form-based searching is perhaps the most useful of our sorting options in that, although tutors can easily select texts on a specific topic or of an appropriate level of difficulty, once they are familiar with appropriate sources from where to take them, it is a very difficult and time-consuming task to find texts that contain specific grammatical forms. Our tools make this task considerably easier and afford tutors access to texts that they can give students as supplementary reading materials to accompany in-class grammar drills. In the case of business Russian or official written discourse, students may extract texts that contain grammatical constructions that occur frequently such as gerunds or participles. Such searches can be done with the underlying CWB system but would be time-consuming and users would need to have in-depth knowledge of the search syntax as well as some computational expertise. We provide a function in the interface that hides this complexity in a manner similar to the way in which we made affix searching more accessible.

10.4 Corpus-Based Exercises

So far in this chapter we have dealt only with corpus-based language learning and teaching from a technological perspective, describing how the tools and functions that we implemented on IntelliText can be used in LSP teaching. However, for these tools to be used effectively, pedagogically relevant and needs-driven materials and exercises need to be written to accompany the technology and a suitable methodology needs to developed for introducing corpora to existing teaching syllabi. Unfortunately, the pedagogical aspect of corpus-based learning and teaching has not been researched in equal measure to the technological side of the approach, and thus material design lags behind advances in the technology, and this is perhaps a reason that DDL has not reached the mainstream since the innovative work of Tim Johns in the 1980s and 1990s (1991a, b, 1993).

We have been developing corpus-based materials on two small projects: ReadingCorp, mentioned above (see Sect. 10.3.2.1), and WritingCorp – Corpus-based exercises for essay and précis writing in Russian – funded by the Teaching Enhancement and Student Success (TESS) fund at the University of Leeds. On ReadingCorp we are using a vocabulary-oriented approach and designing materials on the basis of frequency lists taken from texts that are directly relevant to the research topics of PhD students at the universities of Leeds and Sheffield. Our aim is to use these materials to equip students with skills for comprehending essential primary materials written in Russian. On WritingCorp we are designing corpus-based exercises on themed topics that students cover at Level 2 (crime, health, culture, society, etc.) to allow them to consolidate and expand on material covered in class and improve their written Russian by engaging with topic-related texts.

Tutors can make use of different types of corpus-based exercise. We either create exercises around the results of corpus searches and embed them in a broader pedagogical context or, to encourage learner autonomy, we design exercises that require students to actively work with corpora and perform their own searches, either in language lab sessions or as homework assignments. The former type of exercise has been termed the "hands-off" or "soft" approach to corpus-based language learning and the latter the "hands-on" or "hard" approach by Boulton (2008) and Gabrielatos (2005), respectively. By employing the hands-on or hard approach we not only look to enhance students' language competence but we also teach the students how to use corpora in learning a foreign language – a useful transferable skill that students can apply when learning other languages and an exercise that helps us in our endeavour to foster a culture of corpus use. The hands-on approach is particularly useful from a motivational perspective in that it allows students to work independently outside the classroom and gives them an opportunity to explore language-related issues that are of particular interest to them. Most in-class exercises are centred round the hands-off or soft approach and can be created on the fly for a specific lesson. We also provide hands-off exercises on our VLE system, either as Word documents or as interactive exercises created in Hot Potatoes. In sum, using a corpus-based approach, we aim not only to improve students' language competence but we also train them in using corpora to address language variation and other linguistic issues – a valuable skill that they can apply in other language modules and for languages other than Russian.

We introduce corpora and corpus-based exercises to existing modes of delivery. We believe that corpora should be used to support, not replace, existing teaching practices and we strive towards a blended learning approach, which is highly desirable in the modern-day language learning classroom. We use several types of exercise on the business Russian module, including: (1) a four-step approach to vocabulary building; (2) compound searching; (3) using collocations to find differences between near synonyms and problematic pairs of words; and (4) indentifying the more frequent of two or more expressions with similar usage.

In the first of these exercises, we give students a corpus of texts and ask them to perform several tasks. First, they generate a keyword list. Second, they perform a concordance search on the keywords to highlight their use in a broader context. Third, they perform a collocation search on the keywords to extract common collocations. Finally, they enter the collocations into a concordance search. Working with authentic data, they inevitably find other interesting cases of language use in carrying out these exercises.

For compound searching we identify potentially interesting compound forms, either the initial or end part of a compound form, and ask students to find words that contain these forms and then build on their initial search in various ways. Sample compound forms are given in Table 10.2.

The corpus generates business-related words such as *торгово-промышленный* "commercial and industrial", *налогооблагаемый* "taxable" and *агрофирма* "(an) agro-industrial firm". These words are of little use to students out of context, with respect to active vocabulary acquisition; therefore, we use the hands-on approach

t2.1 **Table 10.2** Sample "parts"
t2.2 of business-related compound
t2.3 forms

Search for as "prefix"	Search for as "suffix"	
работо		t2.5
торгово		t2.6
валютно		t2.7
налого		t2.8
финансово		t2.9
	торговля	t2.10
	фирма	t2.11
	общество	t2.12
	производство	t2.13

titleid	left	match	right
>>	документами банка. — Является ли курсовая разница	налогооблагаемым доходом	? — Курсовая разница в целях налогового учета
>>	на доходы. Еще одна приятность: из вашего	налогооблагаемого дохода	вычитается миллион рублей и проценты по
>>	системе, то есть зависеть от получаемого	налогооблагаемого дохода	. В ту же инспекцию в срок до конца января
>>	, то у такого физического лица возникает	налогооблагаемый доход	. Таки образом, уплата страховых взносов
>>	пенсионные вкладов работника вычитаются из	налогооблагаемого дохода	. После выхода на пенсию, когда накопленные
>>	ведение бизнеса, которые в США вычитаются из	налогооблагаемых доходов	). После появления первой аналоговой копии
>>	фирма не несет никаких расходов, уменьшить	налогооблагаемые доходы	она не вправе. Однако такая позиция контролирующих
>>	подтверждающим документам включается в его	налогооблагаемый доход	и подлежит обложению налогом с доходов физлиц
>>	earnings: совокупный доход. Персональный	налогооблагаемый доход	до проведения всех корректировок. После
>>	разрешалось максимально вычитать из своего	налогооблагаемого дохода	суммы до 40 000 ф. ст., а максимальный капитал
>>	индекс налога и цен): Степень повышения	налогооблагаемого дохода	, необходимая для компенсирования налогоплательщикам
>>	ступени налоговой шкалы): Предельные размеры	налогооблагаемого дохода	или богатства, в границах которых доход
>>	Налоговые декларации содержат сведения о	налогооблагаемых доходах	, таможенные декларации - о провозимых через
>>	предыдущие или последующие годы, чтобы уменьшить	налогооблагаемый доход	соответствующего года. Покупатель. (-0-
>>	прямые налоги, взимаемые в зависимости от	налогооблагаемого дохода	данного лица с учетом предоставленных ему

Fig. 10.8 Concordances generated from the collocation *налогооблагаемый доход*

and instruct them to perform concordance and collocation searches on some of the compound forms and submit their annotated sample sentences to us. A collocation search reveals that *торгово-промышленный* is used almost exclusively in the phrase *торгово-промышленная палата* "Chamber of Commerce and Industry" and that common collocations that contain *налогооблагаемый* are *налогооблагаемый доход* and *налогооблагаемая прибыль*, both meaning "taxable income". Students can then perform a concordance search on these collocations, by clicking on the collocation in the results table, to see them in a broader context. This type of task is particularly suited to homework assignments (Fig. 10.8).

By performing a collocation search on *налогооблагаемый* and then clicking on the concordances of *налогооблагаемый доход* students may also identify other words or phrases that are commonly found in its immediate context: *вычитать (из)* "to deduct (from)", *уменьшить* "to reduce", *предельные размеры* "threshold", etc. Therefore, from the initial word-forming exercise students would be able to commit words like *налогооблагаемый* to their passive vocabulary, while by performing collocation and concordance searches they are more likely to be able to apply them actively by using corpora to see how a word like *налогооблагаемый* is used both in specific collocations and in a broader context.

Corpora are particularly useful for understanding subtle differences in meaning and making distinctions between near synonyms or pairs of words that non-Russians

Pair	Count	F1	F2	LL	MI	T	Pair	Count	F1	F2	LL	MI	T
свой ремесло	134	644414	2003	146	4.48	11.06	розничный торговля	1079	4062	17990	3904.63	11.63	32.84
заниматься ремесло	33	45789	2003	55.56	6.27	5.67	внешний торговля	1029	27249	17990	2654.69	8.81	32.01
художественный ремесло	22	12017	2003	47.18	7.61	4.67	международный торговля	612	35306	17990	1334.56	7.69	24.62
этот ремесло	74	944692	2003	46.63	3.07	7.58	оптовый торговля	396	2358	17990	1327.06	10.97	19.89
народный ремесло	20	18268	2003	37.78	6.87	4.43	в торговля	821	426615	17990	902.47	4.52	27.4
какому-нибудь ремесло	11	559	2003	36.66	11.04	3.32	свободный торговля	414	28512	17990	864.87	7.44	20.23
писательский ремесло	11	1112	2003	32.86	10.05	3.31	биржевой торговля	218	2846	17990	639.08	9.83	14.75
столярный ремесло	8	217	2003	29.21	11.95	2.83	мировой торговля	246	29268	17990	446.22	6.65	15.53
и ремесло	141	5535765	2003	27.11	1.45	7.52	электронный торговля	189	16208	17990	373.51	7.12	13.65
традиционный ремесло	14	13226	2003	26.2	6.82	3.71	министерство торговля	170	14943	17990	333.83	7.08	12.94
актерский ремесло	9	1110	2003	25.98	9.76	3	беспошлинный торговля	69	225	17990	254.54	11.84	8.3
военный ремесло	18	41828	2003	25.71	5.53	4.15	сфера торговля	134	25538	17990	211.58	5.97	11.39
учиться ремесло	14	22040	2003	22.67	6.09	3.69	заниматься торговля	145	45789	17990	193.04	5.24	11.72
кузнечный ремесло	6	204	2003	21.21	11.62	2.45	организация торговля	156	69684	17990	181.47	4.74	12.02
развитие ремесло	18	75399	2003	20.56	4.68	4.08	биржевый торговля	57	585	17990	174.14	10.18	7.54

Fig. 10.9 Collocation-based comparison of the words *ремесло* and *торговля*

this figure will be printed in b/w

find hard to distinguish and for which conventional dictionaries and grammars offer little help. Some typical and more general examples are "crime" and "bed": *преступление* vs. *преступность* and *постель* vs. *кровать*. By viewing the collocates of these words students can achieve a better understanding of how they are used. They should be able to identify that *преступление* denotes a specific crime, whereas *преступность* is an abstract noun denoting crime in general. Likewise, by looking at the adjective collocates on the left of *постель* and *кровать* students should be able to identify that *постель* is a bed in both the physical and abstract sense (as well as bedding), while *кровать* is only a physical object or a bed frame. They can also use corpora to establish important distinctions between problematic pairs of phrases such as *в русском языке* vs. *на русском языке* "in (the) Russian (language)", *в автобусе* vs. *на автобусе* "in/on/by (the) bus) and *в самом деле* vs. *на самом деле* "in actual fact", which many learners of Russian fail to master. In a business context, words such as *business*, *trade* and *commerce* may cause students problems, as each has several translations in Russian. Let us look at *trade* to highlight this point (Fig. 10.9).

The results show that *ремесло* is a specific trade and *торговля* is an abstract term denoting the exchange of goods. Moreover, there are some interesting findings. Both words occur with *заниматься* "to do; be engaged in" and *военное* (*военный* "military") *ремесло* is potentially confusing, as it may be understood by some native speakers of English as meaning "arms trade". In the first case, *заниматься ремеслом* means "to ply a trade", while *заниматься торговлей* means "to be involved in trade; to work as/be a trader (of)". A concordance search of *военное ремесло* would show students that this term denotes the act of training to become a soldier and is therefore a specific trade, as we would expect; "arms trade" is *торговля оружием*, again with the expected Russian equivalent of trade (*торговля*). Therefore, by using a collocation search, if necessary in combination with viewing concordances, students can work out problematic distinctions between words and phrases that cause learners problems more effectively than by relying on (brief) descriptions offered in language manuals. Corpora also highlight anomalous cases in which actual language use does not correspond to the prescribed rules.

Fig. 10.10 Table of results for a frequency comparison of *международная фирма* (Query 1) and *международная компания* (Query 2)

Corpus	Query 1		Query 2	
	Instances	IPM	Instances	IPM
RNC2009-MOCKY	7	0.060	43	0.368
RNC2009-FULL	7	0.060	43	0.369
RS-MOCKY	0	0.000	1	0.172
NEWS-RU	10	0.129	82	1.056
I-RU	11	0.055	96	0.484
BIZ-RU	1	0.063	51	3.205
WIKI-RU	8	0.045	122	0.682
RU-SPOKEN	0	0.000	0	0.000
Totals	44	0.062	438	0.616

Frequency comparisons are useful for identifying the more common of two or more words or phrases that are semantically synonymous, as in: preposition + snail/postal/paper + mail. It is often the case that one of two or more words or phrases that differ neither in meaning (nor in register) is used considerably more. Tutors can give their students sets of examples to compare, and a subsidiary aim of this task is to improve students' analytical skills and for them to make such queries themselves on other pairs of words that they encounter in their own reading.

Two business-related examples are *международная фирма* vs. *международная компания* "international firm, company" and деловая женщина vs. бизнесменка "business woman". In the first case, *фирма* and *компания* are contextually synonymous; in the second, *бизнесменка* is a neologism and considered a colloquial form. A frequency comparison search will show only differences in the frequency of the words or phrases that are compared, not subtle differences in their use. Therefore, to make sure that words are in fact synonyms students are advised to perform a concordance search on each of them. Figures 10.10 and 10.11 show the results for a search performed on *международная фирма* vs. *международная компания*. Students can see from these results that *международная компания* is the more common of the two by a considerable margin – a perhaps surprising observation for learners of Russian. Frequency comparisons may also be used to compare (across different corpora) the use of words or phrases that are semantically synonymous but differ stylistically. A good example is *выйти из дома* vs. *выйти из дому* "to leave the house". The former is neutral, while the latter is colloquial; therefore, students can look at their use in corpora of various genres to achieve a better understanding of when it is appropriate to use one or the other form.

10.5 Conclusions

We have shown in this chapter several ways in which corpora can facilitate and enhance learning and teaching and that a corpus-based approach to business Russian, as well as to LSP teaching more generally, has many benefits and helps to

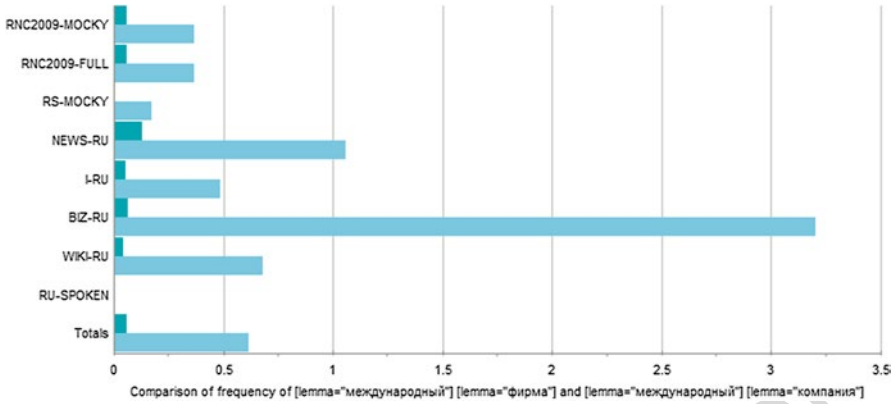


Fig. 10.11 Chart of a frequency comparison search of *международная фирма* (Query 1) and *международная компания* (Query 2)

compensate for the lack of printed teaching materials. We have also shown that corpora are well suited to training certain important LSP skills, such as vocabulary acquisition and register recognition and differentiation.

Corpus-derived frequency lists provide the basis for LSP courses and they can be created for most topics, provided that a large enough body of literature exists for these topics; importantly, a corpus-based approach can meet the needs of the individual language learner, such as a PhD student needing to acquire specialised reading skills in a short time. The breadth of LSP subjects cannot be covered by conventional printed course books, printed course books cannot cater for the needs of the individual and nor can they represent events taking place at the present time. Corpora can cater for all of these needs. Tutors, or even students, can use our tools to generate frequency, collocation and concordance lists, on the basis of which ad-hoc materials and exercises can be created. The materials (and the lists) can be stored electronically, in online repositories, and made available to other users; moreover, available in the format of a collaborative wiki, they can be updated and augmented at any time by their creators or by others. Collaborative learning is particularly important within ICT (Information and Communication(s) Technology) and the pooling and sharing of language resources has many benefits. First, collaboration can enhance language learning and teaching and facilitate tutors' work. Second, a collaborative approach can help provide support both for LSP or other specialised modules and for less-commonly taught languages for which student numbers may not warrant a taught programme but for which tuition can be offered cross-institutionally through a collaborative distance e-learning programme. A further advantage of corpora over printed language manuals is that they are not limited by space. Students can view hundreds of collocations or concordances as opposed to just one or two. A corpus-based approach to LSP teaching can also be used to facilitate register recognition and differentiation. We were able to utilise tools for automatic genre classification to tag business clichés and formal

expressions to help students extract set phrases and use them in their own sample official letters. Other tags can be used in other LSP domains to help learners differentiate between genres and registers more accurately.

That said, for corpora to become a standard feature in LSP teaching more research is needed on both the technological and pedagogical strands of corpus-based learning and teaching. Corpus-based tools require refinement for pedagogical purposes and a more accurate and sophisticated annotation system needs to be developed. Such tools are being developed, yet they are still in their infancy and are in need of further refinement. With regard to pedagogy, relevant and needs-driven materials need to be developed, as does an extensible methodology for the application of corpora in learning and teaching. Without materials to support it the technology is of little practical use to the majority of language learners. More collaboration between computer scientists and researchers in other fields is needed to identify further tools and functions that language learners require. However, it is fair to say on the basis of existing evidence and expected technological advances that corpus-based tools will become easier to use and more functional, tagging will become more accurate and corpora will play an increasingly important role in research and teaching in various academic disciplines, including LSP teaching.

In sum, we believe that a corpus-based approach can make an important contribution to LSP teaching; corpora can facilitate the work of both tutors and their students and they can be used to create materials for highly specialised subjects, on current affairs and to meet the needs of the individual learner. However, specific conditions need to be met for a corpus-based approach to reach the mainstream. First, corpora should be used to support, not replace, existing teaching practices and should be incorporated into a blended learning package optimal for enhancing learners' language competence. Second, corpus-based exercises need to be created to accompany the technology. Third, more work is needed to promote, through workshops, seminars and practical demonstrations, corpus-based language learning in order to present its benefits and methods of application to tutors and students and to foster a culture of corpus use among the language learning and teaching community.

References

Adolphs, S. 2006. *Introducing electronic text analysis*. London: Routledge.

Aijmer, K. (ed.). 2009. *Corpora and language teaching*. Amsterdam: John Benjamins.

Anderson, W., and J. Corbett. 2009. *Exploring English with online corpora*. Basingstoke: Palgrave MacMillan.

Aston, G., S. Bernardini, and D. Stewart (eds.). 2004. *Corpora and language learners*. Amsterdam/ Philadelphia: John Benjamins.

Baroni, M., and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC2004*, Lisbon.

Bennet, G. 2010. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Michigan: University of Michigan Press.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., et al. 1999. *Longman grammar of spoken and written English*. London: Longman.

[AU6] Boulton, A. 2008. DDL: Reaching the parts other teaching can't reach? In *Proceedings of the 8th teaching and language corpora conference*, ed. A. Frankenberg-Garcia. Lisbon: ISLA.

Butler, C. 1974. *German for chemists. Teaching languages to adults for special purposes*, CILT reports and paper 11, 50–53. London: CILT.

Campoy, M., B. Belles-Fortuno, and L. Gea-Valor (eds.). 2010. *Corpus-based approaches to English language teaching*. London: Continuum.

Dlouhá, O. 1998. *Ruština pro veřejnou zprávu*. Plzeň: Aleš Čeněk.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.

Dynda, A., and E. Dyndová. 1998. *Česko-ruská obchodní korespondence*. Prague: Pragoeduca.

Evert, S., and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting of the association for computational linguistics*, Toulouse.

Frankenberg-Garcia, A., L. Flowerdew, and G. Aston (eds.). 2011. *New trends in corpora and language learning*. London: Continuum.

Gabrielatos, C. 2005. Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ* 8(4): 1–35.

Golčáková, B. 2008. *Ruština pro mezinárodní vztahy*. Plzeň: Aleš Čeněk.

Harris, T., and M. Moreno (eds.). 2010. *Corpus linguistics in language teaching*. Bern: Peter Lang.

Hidalgo, E., L. Quereda, and J. Santana (eds.). 2007. *Corpora in the foreign language classroom*. Amsterdam: Rodopi.

Johns, T. 1991a. Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal* 4: 1–16.

Johns, T. 1991b. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *English Language Research Journal* 4: 27–45.

Johns, T. 1993. Data-driven learning: An update. *TELL&CALL* 2: 4–10.

Kilgarriff, A. 2010. Comparable corpora within and across languages: Word frequency lists and the Kelly project. In *Proceedings of the BUCC workshop*, Malta.

Kilgarriff, A., M. Husák, K. McAdam, M. Rundell, and P. Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX'08*, Barcelona.

Kotani, K., T. Yoshimi, T. Kutsumi, I. Sata, and H. Isahara. 2008. EFL learner reading time model for evaluating reading proficiency. In *Proceedings of CICLING*, Haifa.

Kupcevičová, J., and V. Vilímek. 2006. *Ruská řečová etiketa*. Ostrava: Philosophical Faculty of Ostrava University.

Mason, J., M. Shepherd, J. Duffy, V. Keselj, and C. Watters. 2010. An n-gram based approach to multi-labeled web page genre classification. In *Proceedings of 43rd Hawaii international conference on system sciences*, Hawaii.

McCarthy, M. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41(4): 563–574.

McEnery, T., and A. Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Mrověcová, L. 2007. *Obchodní ruština*. Brno: Computer Press.

O'Keeffe, A., and M. McCarthy (eds.). 2010. *The Routledge handbook of corpus linguistics*. London: Routledge.

Rayson, P., and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the comparing corpora workshop at ACL*, Hong Kong.

Reppen, R. 2010. *Using corpora in the language classroom*. Cambridge/New York: Cambridge University Press.

Rezková, J. 2008. *Ruština pro právníky*. Prague: Univerzita Karlova.

Segler, T. 2007. *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Unpublished Ph.D. thesis, University of Edinburgh.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as corpus*, ed. M. Baroni and S. Bernardini. Bologna: Gedit.

Sharoff, S. 2010. In the garden and in the jungle: Comparing genres in the BNC and internet. In *Genres on the web: Computational models and empirical studies*, ed. A. Mehler, S. Sharoff, and M. Santini. Berlin/New York: Springer.

Sharoff, S., S. Kurella, and A. Hartley. 2008. Seeking needles in the Web haystack: Finding texts suitable for language learners. In *Proceedings of teaching and learning corpora conference*, Lisbon.

Sharoff, S., Z. Wu, and K. Markert. 2010. The Web library of Babel: evaluating genre collections. In: *Proceedings of the seventh language resources and evaluation conference*, Malta.

Sinclair, J. (ed.). 2004. *How to use corpora in language teaching*. Amsterdam: John Benjamins.

Willis, D. (ed.). 2004. *Collins cobuild – Intermediate English grammar*. Birmingham: University of Birmingham.

Wilson, J., A. Hartley, S. Sharoff, and P. Stephenson. 2011. Advanced corpus solutions for humanities researchers. In *Proceedings of Pacific Asia conference on language, information and Computation Sendai*, 2010.

Witten, I., and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann.