

a **FREQUENCY** dictionary of
RUSSIAN

CORE VOCABULARY FOR LEARNERS

Serge Sharoff, Elena Umanskaya
and James Wilson

● Practical: the top 5,000 most frequently used Russian words

● Learner friendly: gives you the core vocabulary for Russian quickly

● Useful: 26 thematic boxes give the top words for specific topics

A Frequency Dictionary of Russian

A Frequency Dictionary of Russian is an invaluable tool for all learners of Russian, providing a list of the 5,000 most frequently used words in the language and the 300 most frequent multiword constructions.

The dictionary is based on data from a 150-million-word Internet corpus taken from more than 75,000 webpages and covering a range of text types from news and journalistic articles, research papers, administrative texts and fiction.

All entries in the rank frequency list feature the English equivalent, a sample sentence with English translation, a part of speech indication, indication of stress for polysyllabic words and information on inflection for irregular forms.

The dictionary also contains twenty-six thematically organized and frequency-ranked lists of words on a variety of topics, such as food and drink, travel, and sports and leisure.

A Frequency Dictionary of Russian enables students of all levels to get the most out of their study of vocabulary in an engaging and efficient way. It is also a rich resource for language teaching, research, curriculum design and materials development.

A CD version is available to purchase separately. Designed for use by corpus and computational linguists it provides the full text in a tab-delimited format that researchers can process and turn into suitable lists for their own research purposes.

Serge Sharoff is Senior Lecturer and **James Wilson** is Research Fellow, both at the Centre for Translation Studies within the School of Modern Languages and Cultures, University of Leeds.

Elena Umanskaya is a freelance teacher of Russian as a foreign language.

Routledge Frequency Dictionaries

General Editors

Paul Rayson, *Lancaster University, UK*

Mark Davies, *Brigham Young University, USA*

Editorial Board

Michael Barlow, *University of Auckland, New Zealand*

Geoffrey Leech, *Lancaster University, UK*

Barbara Lewandowska-Tomaszczyk, *University of Lodz, Poland*

Josef Schmied, *Chemnitz University of Technology, Germany*

Andrew Wilson, *Lancaster University, UK*

Adam Kilgarriff, *Lexicography MasterClass Ltd and University of Sussex, UK*

Hongying Tao, *University of California at Los Angeles*

Chris Tribble, *King's College London, UK*

Other books in the series

A Frequency Dictionary of Arabic

A Frequency Dictionary of Czech

A Frequency Dictionary of Contemporary American English

A Frequency Dictionary of Dutch (forthcoming)

A Frequency Dictionary of German

A Frequency Dictionary of French

A Frequency Dictionary of Japanese

A Frequency Dictionary of Mandarin Chinese

A Frequency Dictionary of Portuguese

A Frequency Dictionary of Spanish

A Frequency Dictionary of Russian

Core vocabulary for learners

Serge Sharoff, Elena Umanskaya and James Wilson

First published 2013

by Routledge

2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada

by Routledge

711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2013 Serge Sharoff, Elena Umanskaya and James Wilson

The right of Serge Sharoff, Elena Umanskaya and James Wilson to be identified as authors of this work has been asserted by them in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

Sharoff, Serge.

A frequency dictionary of Russian : core vocabulary for learners / Serge Sharoff, Elena Umanskaya and James Wilson.

pages ; cm. – (Routledge frequency dictionaries)

Includes bibliographical references and index.

1. Russian language – Word frequency – Dictionaries. I. Umanskaya, Elena.
II. Wilson, James, 1979 – III. Title. IV. Series: Routledge frequency dictionaries.
PG2691.S44 2013
491.73'21 – dc23

2012031935

ISBN: 978-0-415-52141-3 (hbk)

ISBN: 978-0-415-52142-0 (pbk)

ISBN: 978-0-415-66305-2 (CD)

Typeset in Parisine and Arial

by Graphicraft Limited, Hong Kong



Printed and bound in Great Britain by
TJ International Ltd, Padstow, Cornwall

Contents

Thematic vocabulary lists | **vi**

Series preface | **vii**

Acknowledgements | **ix**

Abbreviations | **x**

Introduction | **1**

Frequency index | **13**

Alphabetical index | **278**

Part-of-speech index | **331**

Most common multiword expressions | **369**

Thematic vocabulary lists

- 1 Animals | **22**
- 2 Clothing | **32**
- 3 Colours | **42**
- 4 Communication | **52**
- 5 Directions and location | **62**
- 6 Drinks | **72**
- 7 Expressing motion | **82**
- 8 Food | **92**
- 9 Friends and family | **102**
- 10 Fruit and vegetables | **112**
- 11 Health and medicine | **122**
- 12 House and home | **132**
- 13 Human body | **142**
- 14 Language learning | **152**
- 15 Moods and emotions | **162**
- 16 Numbers | **172**
- 17 Popular festivals | **182**
- 18 Professions | **192**
- 19 School and education | **202**
- 20 Size and dimensions | **212**
- 21 Sports and leisure | **222**
- 22 The natural world | **232**
- 23 Time expressions | **242**
- 24 Town and city | **252**
- 25 Travel | **262**
- 26 Weather | **272**

Series preface

Frequency information has a central role to play in learning a language. Nation (1990) showed that the 4,000–5,000 most frequent words account for up to 95 per cent of a written text and the 1,000 most frequent words account for 85 per cent of speech. Although Nation's results were only for English, they do provide clear evidence that, when employing frequency as a general guide for vocabulary learning, it is possible to acquire a lexicon which will serve a learner well most of the time. There are two caveats to bear in mind here. First, counting words is not as straightforward as it might seem. Gardner (2007) highlights the problems that multiple word meanings, the presence of multiword items, and grouping words into families or lemmas present in counting and analysing words. Second, frequency data contained in frequency dictionaries should never act as the only information source to guide a learner. Frequency information is nonetheless a very good starting point, and one which may produce rapid benefits. It therefore seems rational to prioritise learning the words that you are likely to hear and read most often. That is the philosophy behind this series of dictionaries.

Lists of words and their frequencies have long been available for teachers and learners of language. For example, Thorndike (1921, 1932) and Thorndike and Lorge (1944) produced word frequency books with counts of word occurrences in texts used in the education of American children. Michael West's *General Service List of English Words* (1953) was primarily aimed at foreign learners of English. More recently, with the aid of efficient computer software and very large bodies of language data (called corpora), researchers have been able to provide more sophisticated frequency counts from both written text and transcribed speech. One important feature of the resulting frequencies presented in this series is that they are derived from recently collected language data. The earlier lists for English included samples from, for example, Austen's *Pride and Prejudice* and Defoe's *Robinson Crusoe*, thus they could no longer represent present-day language in any sense.

Frequency data derived from a large representative corpus of a language brings students closer to language as it is used in real life as opposed to textbook language (which often distorts the frequencies of features in a language, see Ljung, 1990). The information in these dictionaries is presented in a number of formats to allow users to access the data in different ways. So, for example, if you would prefer not to simply drill down through the word frequency list, but would rather focus on verbs for example, the part-of-speech index will allow you to focus on just the most frequent verbs. Given that verbs typically account for 20 per cent of all words in a language, this may be a good strategy. Also, a focus on function words may be equally rewarding – 60 per cent of speech in English is composed of a mere 50 function words. The series also provides information of use to the language teacher. The idea that frequency information may have a role to play in syllabus design is not new (see, for example, Sinclair and Renouf, 1988). However, to date it has been difficult for those teaching languages other than English to use frequency information in syllabus design because of a lack of data.

Frequency information should not be studied to the exclusion of other contextual and situational knowledge about language use and we may even doubt the validity of frequency information derived from large corpora. It is interesting to note that Alderson (2007) found that corpus frequencies may not match a native speaker's intuition about estimates of word frequency and that a set of estimates of word frequencies collected from language experts varied widely. Thus corpus-derived frequencies are still the best current estimate of a word's importance that a learner will come across. Around the time of the construction of the first machine-readable corpora, Halliday (1971: 344) stated that "a rough indication of frequencies is often just what is needed". Our aim in this series is to provide as accurate as possible estimates of word frequencies.

Paul Rayson and Mark Davies
Lancaster and Provo, 2008

References

- Alderson, J. C. (2007)** "Judging the frequency of English words." *Applied Linguistics*, 28 (3): 383–409.
- Gardner, D. (2007)** "Validating the construct of Word in applied corpus-based vocabulary research: a critical survey." *Applied Linguistics*, 28, pp. 241–65.
- Halliday, M. A. K. (1971)** "Linguistic functions and literary style." In S. Chatman (ed.) *Style: A Symposium*. Oxford University Press, pp. 330–65.
- Ljung, M. (1990)** *A Study of TEFL Vocabulary*. Almqvist & Wiksell International, Stockholm.
- Nation, I. S. P. (1990)** *Teaching and Learning Vocabulary*. Heinle & Heinle, Boston.
- Sinclair, J. M. and Renouf, A. (1988)** "A lexical syllabus for language learning." In R. Carter and M. McCarthy (eds) *Vocabulary and Language Teaching*. Longman, London, pp. 140–58.
- Thorndike, E. (1921)** *Teacher's Word Book*. Columbia Teachers College, New York.
- Thorndike, E. (1932)** *A Teacher's Word Book of 20,000 Words*. Columbia University Press, New York.
- Thorndike, E. and Lorge, I. (1944)** *The Teacher's Word Book of 30,000 Words*. Columbia University Press, New York.
- West, M. (1953)** *A General Service List of English Words*. Longman, London.

Acknowledgements

The development of the corpus and the tools for processing Russian received funding from EPSRC grant EP/C005902 (Project ASSIST), and the EU FP7 programme under Grant Agreement No 248005 (Project TTC). The initial stage for preparation of the frequency lists received funding from the EU LLP-KA2 Programme, 505630-LLP-1-2009-1-SE-KA2-KA2MP (Project Kelly).

Abbreviations

The following part-of-speech codes have been used:

A	Adjective
Adv	Adverb
C	Conjunction
I	Interjection
Nc	Noun, common gender, e.g., убийца , killer, which can be used as either a masculine or feminine noun in the same form
Nf	Noun, feminine
Nm	Noun, masculine
Nn	Noun, neutral
N-	Noun (existing in plural form only, so no gender can be indicated)
Num	Numeral
P	Pronoun
Part	Participle
Prep	Preposition
V	Verb

Introduction

1 Corpora and frequency lists for language learners

Corpus-based approaches to defining the language curriculum are not new. The assumption that more common words are more useful to language learners has been tested in various studies, starting with the works of Michael West in the 1930s on the *General Service List* (West, 1953) and by Thorndike and Lorge on the *Teacher's Word Book* (Thorndike and Lorge, 1944). Developments in the field of computer technology led to the proliferation of statistical studies of word frequency from the 1960s (Juilland, 1964; Kučera and Francis, 1967; Juilland et al., 1970), and frequency dictionaries for Russian were developed around this time as well (Shteynfel'd, 1963; Zazorina, 1977).

Corpus-derived frequency lists are based on objective word counts; that is, words are 'arranged according to the number of times they occur in particular samples of language' (Richards, 1974: 71). The pedagogical relevance of such word lists has been brought into question in that (1) lists differ, sometimes quite substantially, depending on their source (i.e. the corpus from which they were extracted), and (2) many common words are often absent from such lists. With regard to point (2), words like *soap*, *soup*, *bath* and *trousers* do not appear in the first 2,000 words of a 30,000-word frequency list compiled by Thorndike and Lorge (1944); likewise, in other frequency lists compiled by Earnest Horn, John Dewey and Edward Thorndike words like *dispose*, *err* and *execute* appeared among the first 1,000, while *animal*, *hungry* and *soft* did not.

Gougenheim et al. (1956) were probably the first to notice that 'objective' frequency lists lack some everyday words (*mots disponibles*), which most speakers of a language would consider common. This problem was referred to as the problem of *oranges and bananas* in the Kelly project (Kilgarriff, 2010), because traditional corpora often lack words of this sort. For this reason, the relationship between the frequency of words and their pedagogical

relevance has been questioned, and many researchers believe that word frequency is too problematic to be useful.

Nevertheless, corpus-derived frequency data are invaluable for syllabus and materials design, as evident from the success of this current series (Xiao et al., 2009; Cermák and Kren, 2010; Davies and Gardner, 2010).

Language teachers know intuitively what is suitable for learners, but frequency lists can both support and challenge their intuitions (Alderson, 2007). Pedagogic studies demonstrate the relevance of using frequency lists in language teaching (Bauer and Nation, 1993; Nation, 2004). Extracting frequency lists from corpora is now a standard practice in many areas of lexicography and many modern dictionaries and, increasingly, grammars are corpus-based. Kilgarriff (2010) writes that there are three methods of producing word frequency lists: by (1) copying, (2) guessing and (3) counting (i.e. from corpora); he goes on to state that now corpora are available for many languages, the 'corpus' approach *must* be used. Corpus research has an important role in defining teaching curricula because corpus data show 'which language items and processes are most likely to be encountered by language users, and which therefore may deserve more investment of time in instruction' (Kennedy, 1998: 281). Römer (2008: 115) writes that while word frequency is not the only criterion that should inform decisions regarding the inclusion of words in teaching programmes and curricula, it as an 'immensely important one'; a similar view is expressed by Leech (1997: 16) and Aston (2000: 8).

Moreover, some of the problems outlined above may be linked to limitations in technology and/or available corpora. A corpus is only as good as its contents and the same holds for frequency lists. Nowadays, corpora are much larger (some are made up of hundreds of millions or even more than a billion words), balanced and built to be representative of a language variety; therefore, the

results obtained from these corpora are more 'reliable'. Since the earlier studies mentioned above were published more texts have become available in electronic form and computing power is much greater, making it easier to collect large corpora and produce more reliable frequency lists, e.g. for English (Leech et al., 2001; Davies and Gardner, 2010). Yet there are, of course, still anomalies: some frequent words do not show up in frequency lists, while some obscure or domain-specific words do. A way of overcoming this problem is manually to 'clean' the lists. Waddington (1998) argues that words in frequency lists need to be checked against 'commonsense observations'; tutors may thus review and fix any problems by taking out anomalous words or adding any common words that for whatever reason were absent from the original list. This method was used on the Kelly project (Kilgarriff, 2010), on which cleaned corpus-derived frequency lists for nine languages (Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish), each of 9,000 words, were created. The cleaned list of Russian words developed in Kelly served as the basis for the list of words presented in this dictionary.

Tutors may introduce frequency list data to their students in numerous ways, or students may use frequency lists to structure their own language learning. Tutors may test students on the lists to monitor their progress in vocabulary acquisition – such an approach is especially useful at the *ab initio* level – or they may incorporate the words in language-learning exercises and teaching materials. Students may work through the lists systematically and test themselves at regular intervals or they may use them for reference as a guide to their progress. While grammar is considered by many learners to be the hardest part of learning Russian, there is a finite number of rules and forms that can be taught systematically. Vocabulary, on the other hand, is much harder to teach in a structured way, as there are thousands of words in a language and it is difficult to know which of these words should be introduced to students and when. Brown (1996: 2) writes that 2,000 words may be considered a core vocabulary for a British A-Level Russian language course and the recognition of 2,000 words guarantees at least 75 per cent of the words in a Russian text; he considers 8,000 words, guaranteeing

the recognition of over 90 per cent of words in any Russian text, the target for a university graduate. He writes: 'Any foreign student with a sound knowledge of Russian grammar and a passive knowledge of 8,000 to 10,000 vocabulary items (with perhaps an active vocabulary of half that) can reasonably call him or herself competent in the language for all normal purposes.' Word frequency lists, especially those annotated and adapted for language-learning purposes, support vocabulary acquisition by informing teachers and students of the most common words in a language, and allow them to structure the teaching or learning of vocabulary more effectively. They may be used *indirectly* in materials or syllabus design or applied *directly* in the classroom and integrated among core learning activities and/or used for independent self-study and progress monitoring.

2 The Russian Internet Corpus

The dictionary is based on the Russian Internet Corpus, I-RU (Sharoff, 2006), which consists of more than 150 million orthographic words taken from more than 30,000 webpages. More precisely, it contains 198,509,029 tokens (counting orthographic words, numbers and punctuation marks), 159,175,960 words (including words written in both Cyrillic and Latin characters) or 147,803,971 words consisting entirely of Cyrillic characters. The corpus was collected in 2005 according to a method of making queries to Google and collecting the top ten pages retrieved for each query. Although we may question the quality of texts available on the web, a closer investigation of this corpus (Sharoff, 2006; Sharoff, 2007) shows that the Internet does not consist of 'porn and spam'. Traditional corpora like the British National Corpus (Aston, 2000) or the Russian National Corpus (Sharoff, 2005) have been collected manually. Therefore, it is possible to describe the properties of their documents manually as well. Manual annotation is not feasible for a corpus of 30,000 pages, so we have attempted to estimate its contents in two ways.

An automated estimate of the genre composition of I-RU given in Table 1 is based on supervised machine learning. The computer learns statistically significant features of texts belonging to known genre categories to recognize texts in the corpus. The accuracy of machine learning in this task is

Table 1 Genres of I-RU

Genre	Percentage
Reporting (newswires)	10.24%
Fiction and popular lore	27.46%
Legal texts	0.07%
Instruction (FAQ&teaching)	6.88%
'Discussion' (argumentative texts)	55.12%

about 70 to 75 per cent (Sharoff, 2010), so we need to treat the accuracy of each individual figure with caution. Nevertheless, this method gives us a useful estimate of the distribution of genre categories found in the corpus and in the Russian Internet overall. It is known that fiction is under-represented on the web for many languages (Sharoff, 2006), but for Russian the situation is different: a considerable amount of modern fiction is available, and the unclear copyright status of fiction produced during the Soviet era means that it is available as well. Thus, the Russian Internet may be seen as representative of what the Russian population reads at the moment. The largest category of 'Discussion' contains various argumentative texts, including newspaper opinion texts, research papers, student essays, forums and blogs, etc.

Another way of approximating the composition of the Russian Internet corpus is by arranging its documents in a number of dimensions according to their internal similarity to known texts (Forsyth and Sharoff, 2011). We rated eighty-seven documents according to seventeen textual parameters such as:

Argumentative To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?

Instructive To what extent does the aim of the text seem to be to teach the reader how to do something (e.g. a tutorial)?

Promotional/commercial To what extent does the document promote a commercial product or service?

Then we merged the scores into the two most significant dimensions using multi-dimensional scaling (Sammon, 1969) and applied Machine

Learning (SVM regression) to estimate the position of all texts in this corpus. We also applied the same procedure to texts from known categories, which were selected as representing the Brown Corpus categories (Kučera and Francis, 1967): A (news), B (editorials), C (reviews), down to categories K–R (different kinds of fiction). The heatmap in Fig. 1 shows that the most frequent text types approximate fiction, fiction-like texts in the upper-right corner (often they are personal blogs), and news texts on the left side of the picture, extending from news (Category A in the bottom) to editorial-like argumentative texts (Category B).

An interesting issue for language learning concerns the overall size of the lexicon and the portion of the lexicon needed for learners. The total number of orthographic Cyrillic-only lemmas in the lexicon of this corpus is 1,078,346 (after unification for the lower- and upper-case characters); however, only 513,184 of them occur in this corpus more than once: 154,890 lemmas occur more than ten times. The total number of Cyrillic word forms was 1,900,791, while the number of Cyrillic word forms occurring more than ten times was 405,635. In spite of the fact that Russian is considered to be a morphologically rich language, the ratio of forms to lemmas in the entire corpus appears to be relatively small: 1.76 forms per lemma. However, if we take into account only the words occurring in the dictionary, the ratio raises to 8.35 (41,729 attested forms for the 5,000 lemmas), which is a good estimate for the productivity of Russian lemmas. As expected, the verbs (including participles) have the largest number of forms per lemma, 34.56 (32,420 forms per 938 lemmas), with the ratio for the nouns of 8.18 (21,292 attested forms per 2602 lemmas).

Finally, it is possible to estimate the relationship between the lexicon presented in this dictionary and the coverage of texts in the corpus. In Fig. 2 we illustrate the amount of the corpus covered by words up to a given rank. In total, the 5,000 words from this dictionary cover 90.40 per cent of texts in this corpus; the top 2,000 words cover 80 per cent of texts.

3 Existing frequency lists for Russian

As mentioned above, existing lists are outdated and/or not suitable for learners. Frequency dictionaries of Russian appeared fairly early

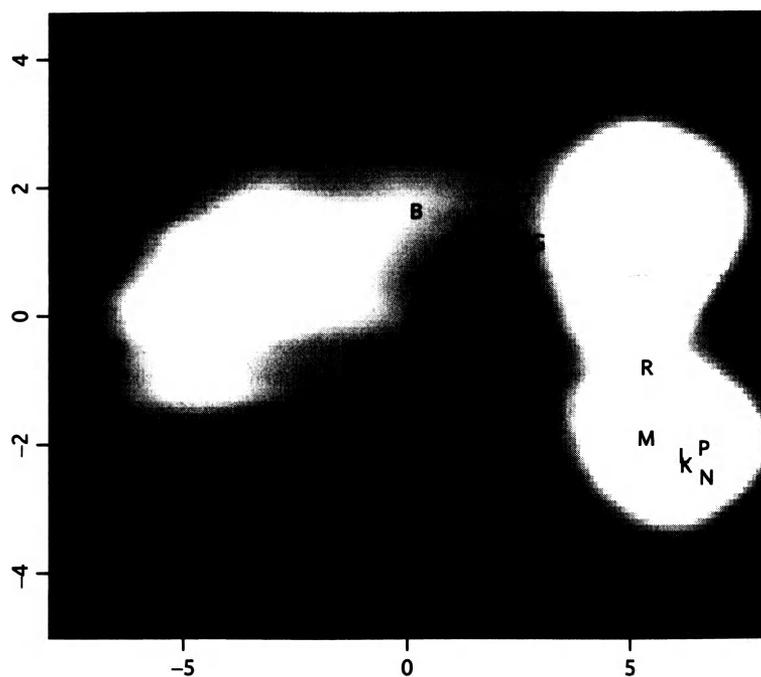


Fig. 1 Distribution of text types in I-RU (i-ru-compos)

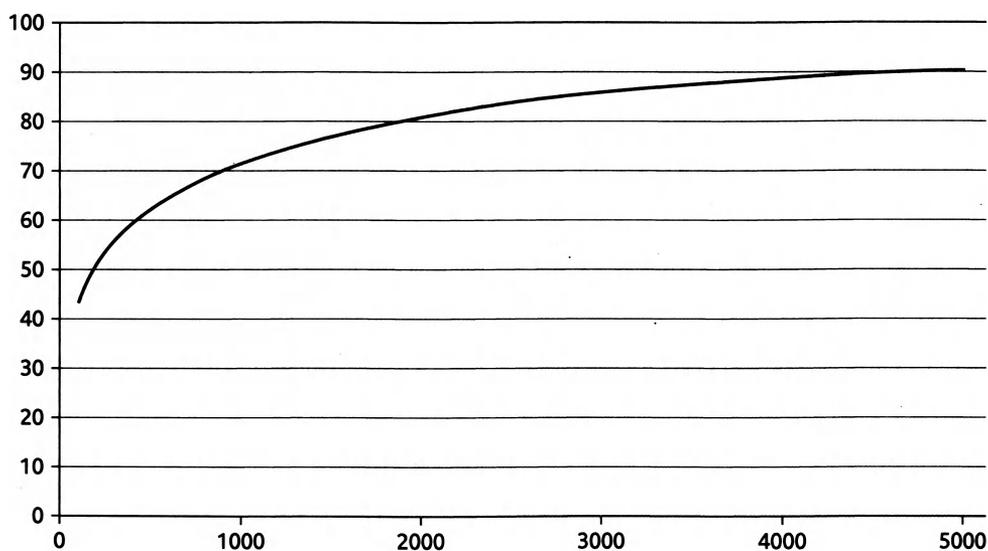


Fig. 2 Lexical coverage

(Shteynfel'd, 1963; Zazorina, 1977), but they were based on relatively small collections of texts; therefore, their word lists are not reliable. Moreover, the sources of these texts from the Soviet era make them seriously outdated now; for example, советский 'Soviet', товарищ 'comrade' and борьба

'struggle' are in the first hundred in the Zazorina list, on a par with function words. The most recent proper frequency list (Lönngrén, 1993) is based on the Uppsala corpus, which is still small by modern standards. It consists of one million words, with an approximately equal amount of fiction and

journalistic texts published between 1960 and 1987. The word list included in Nicholas Brown's *Learner Dictionary* (Brown, 1996) is an adaptation of the Zazorina frequency list produced by moving the Communist vocabulary of Lenin, Khrushchev and Soviet newspapers down the frequency list. However, this dictionary is not a proper frequency dictionary per se; human judgement does not correlate with actual frequencies (Alderson, 2007), while the Zazorina list is based on a very small corpus, so it is not reliable in itself. Brown mentions editing the frequency of **катер** 'boat', but many other words, like **пауза** 'pause' and **молчать** 'keep silence', are also disproportionately more frequent in the Zazorina list.

There is a more modern Russian National Corpus (Sharoff, 2005) containing about 90 million words from a range of sources covering texts from the 1950s to 2000s. The corpus also resulted in a frequency dictionary (Ljashevskaja and Sharoff, 2009), which contains a list of about 50,000 words with information on their frequency distribution by years and genres. However, it is an academic publication with information entirely in Russian and with little potential for its use in foreign language teaching. Besides, even though the RNC is considerably bigger than corpora from which previous Russian frequency lists have been extracted, I-RU is nearly twice the size of the RNC. Table 2 also indicates some of the problems with the RNC frequency list.

Forums and blogs available in I-RU provide an account of the language of personal interaction, which is important to language learners. An example comparing the frequency of some words in I-RU against the Russian National Corpus (RNC) is given in Table 2. Studies of other corpora derived from the Web (e.g. Ferraresi et al., 2008) also show that in comparison to traditional corpora, web corpora

contain more words related to personal interaction, like first- and second-person pronouns and verbs in the present tense.

This stems from the fact that traditional corpora cannot fully represent spontaneous personal interaction. It is quite difficult to collect a sufficient amount of spoken language data, and the compilers had to rely on written sources, while web corpora contain some material (e.g. from blogs) that may be seen as an approximation to the language of personal interaction, and such materials is useful for language learners.

As for domains, I-RU is based on a much larger number of sources than traditional manually collected corpora. It is inevitable that some words become over-represented in traditional corpora, since the amount of sources for each text type is usually limited by what was available to researchers responsible for their collection. Adam Kilgarriff refers to this as a 'whelk problem'; that is, if a text is about *whelks*, the frequency of this word becomes disproportionately high (Kilgarriff, 1997). The RNC contains a number of memoirs of former actors and theatre directors, the business section of the Russian legal code (partly responsible for the frequency of the formal reference to Russia as the *Russian Federation* in Table 2), and a large number of medical texts. The number of different sources of I-RU results in a better coverage of *core* vocabulary, as individual topics of each document are levelled out. Overall, I-RU provides the most reliable frequency list currently available for Russian language learners.

4 Facts about Russian

Russian, or Contemporary Standard Russian (Современный русский литературный язык),

Table 2 Comparing the frequencies in I-RU against RNC (data per million words)

Word	I-RU	RNC	Word	I-RU	RNC
Personal interaction words			Topic-specific words		
я 'I'	10146	9714	премьера 'première'	9	90
ты 'you, fam'	2530	2390	театр 'theatre'	86	303
вы 'you, polite'	2918	2194	арбитражный 'arbitrage'	7	55
спасибо 'thank you'	151	91	Федерация 'Federation'	83	255
пожалуйста 'please'	104	72	вирус 'virus'	20	107
давай / давайте 'let's'	106	84	штамм 'virus strain'	1	36

is a Slavonic language, in the East Slavonic group (together with Ukrainian and Belarusian), spoken as a native language by approximately 150 million people. Russian is the official state language of Russia as well as an official language in Belarus, Kazakhstan, Kyrgyzstan and Tajikistan; it is also widely spoken in other countries of the former USSR as well as in Russian diaspora communities throughout the world.

The Russian (Cyrillic) alphabet is made up of thirty-three letters: twenty-one consonants, ten vowels and the soft (ь) and hard (ъ) signs. A phonological description of Russian is somewhat complicated, as there is disagreement with regard to how many phonemes (the number of distinguishable sounds) make up the Russian sound system. It is generally accepted that Russian has five vowel phonemes (/a/, /e/, /i/, /o/ and /u/), though linguists of the Leningrad School attach phonemic status to */i/ (ы), which is considered an allophone (a variant of a phoneme that occurs only in specific positions) of /i/ (и) by most other linguists. Russian has at least thirty-two consonant phonemes. Moscow School linguistics distinguish thirty-four consonant phonemes and Leningrad School linguists thirty-seven; according to most works in the Western literature on Russian phonology, Russian has either thirty-two or thirty-three consonant phonemes. For a more detailed description of Russian phonology readers are directed to Timberlake (1993: 828–836), Hamilton (1980), and Townsend and Janda (1996: 252–258).

Russian is characterized by mobile stress. Stress in Russian is contrastive and serves to differentiate meaning, either (1) marking differences between words (lexical differences) or (2) marking differences in the grammatical forms of the same word (grammatical differences). For (1), examples such as **замок** 'castle' vs **замок** 'lock', and **мука** 'torment; torture' vs **мука** 'flour' highlight this point; many more heteronyms (words that share the same spelling but have a different pronunciation and meaning) are identified when inflected forms are considered: **белка** 'squirrel' vs **белка** 'egg white' (Gen. Sing.), **ворона** 'raven' (Gen. Sing.) vs **ворона** 'crow', **потом** 'sweat' (Instr. Sing.) vs **потом** 'then, later'. For (2), examples include **города** 'town' (Gen. Sing.) vs **города** 'towns' (Nom./Acc. Pl.), **окна** 'window' (Gen. Sing.) vs **окна** 'windows'

(Nom./Acc. Pl.) and **смотрите** 'you look (watch); you are looking (watching)' (2nd Pers. Pl.; indicative mood) vs **смотрите** 'look, watch' (2nd Pers. Pl.; imperative mood). As stress in Russian is important, stress marks are included in the list of headwords, but stress is not indicated in the examples. Information about stress was taken from the Russian wiktionary.

Russian is a morphologically complex and highly inflected language. Nouns, adjectives and pronouns are inflected according to gender (masculine, feminine and neuter), number (singular and plural) and case (nominative, accusative, genitive, dative, instrumental and prepositional). There is a fairly high level of syncretism between forms across the cases, especially in adjectival and pronominal morphology (and to a lesser degree in nominal morphology). For example, **моей** and **новой** are feminine singular genitive, dative, prepositional and instrumental forms of the pronoun **мой** 'my' and adjective **новый** 'new', respectively; **моих** and **новых** are genitive and prepositional plural forms of these words. Feminine nouns have the same form in the dative and prepositional singular, and neuter nouns of the **время** 'time' type have the same form in the genitive, dative and preposition singular (**времени**). Old Russian had a dual number, but as in many other contemporary Slavonic languages, with the notable exception of Slovene, in modern Russian only vestiges of the dual remain (e.g. **уши** 'ears' or forms that occur after the numeral 2 (and also 3 and 4) that have been re-categorized as the genitive singular (**два часа** 'two hours'), or in many other Slavonic languages replaced by plural forms).

There are also three other cases in Russian: the partitive genitive, the second prepositional (locative) and the vocative. The partitive genitive is used to denote 'a quantity of' and is common with certain verbs (**хотеть** 'to want', **налить** 'to pour', **выпить** 'to drink' as well as with several verbs beginning with the prefix **на-**); masculine nouns have an ending (**сыр** 'cheese' / **сыру**, **чай** 'tea' / **чаю**) distinct from that of the 'regular' genitive (**сыр** / **сыра**, **чай** / **чая**), though the regular forms are increasingly common in partitive genitive contexts, while feminine and neuter nouns have the same ending as in the 'regular' genitive (see Wade 1992: 56 and 89–92 for a more detailed description). The second prepositional or locative case is used to denote

location with the prepositions **в** 'in' and **на** 'in, on'; it does not occur with other prepositions that govern the prepositional case (cf. **в саду** 'in the garden' vs **о саде** 'about the garden'). The vocative case, common to other Slavonic languages such as Bulgarian (in which grammatical case has been lost, barring a few exceptions), Czech and Polish, in Russian is used 'colloquially' in some proper nouns (people's names) and common nouns denoting people (mum, dad, grandma, etc.): word-final consonant phonemes are dropped in mono- and disyllabic words, as in the examples **мама** 'mum' / **мам**, **папа** 'dad' / **пап**, **Таня** 'Tanya' / **Тань** and **Коля** 'Kolya' / **Коль**. It is also used vestigially in religious words: **боже** (from **бог** 'God'), **господи** (from **господь** 'Lord') and **отче** (from **отец** 'father'), as in **Отче наш** 'Our Father' (The Lord's Prayer).

Russian verbal morphology is dominated by verbal aspect. Most Russian verbs have an imperfective and perfective form (e.g. **читать** / **прочитать** 'to read', **объяснять** / **объяснить** 'to explain'); the imperfective form comes before the forward slash. Some verbs are only imperfective (e.g. **наблюдать** 'to observe', **нуждаться** 'to need'), or only perfective (e.g. **очутиться** 'to find oneself', **понадобиться** 'to come in handy'). Some verbs are bi-aspectual (e.g. **исследовать** 'to research', **велеть** 'to command'). Aspectual pairs are formed by: (1) modification to the verbal suffix (e.g. **получать** / **получить** 'to receive'); (2) prefixation (e.g. **смотреть** / **посмотреть** 'to look; watch'); (3) internal modification (e.g. **выбирать** / **выбрать** 'to choose'); in addition, (4) a few verbs have different roots (e.g. **говорить** / **сказать** 'to say', **брать** / **взять** 'to take'). Russian verbs are categorized into finite, infinitive, participle and gerund forms; they have four moods (indicative, conditional, subjunctive and imperative) and three tenses (past, present and future). The past tense has two forms, imperfective and perfective, (past-tense forms of the verb **читать** 'to read', for example, are **читал** (Imperf.) and **прочитал** (Perf.)), as does the future (**буду читать** (Imperf.) and **прочитаю** (Perf.)), while the present tense has just one (**читаю**). Some language tutors try to map Russian aspect to the English tenses, though this is only partially successful. In very simplistic terms, the imperfective is used for durative, habitual, incomplete or unsuccessful actions as well as for general statements; certain verbs also require an

imperfective. The perfective is used for single and completed actions and with certain verbs.

Aspect affects not only the past and future tenses but also infinitives, conditional statements and imperatives. Russian verbs conjugate according to person, tense and mood. Present-tense and perfective future-tense verbs have six forms, as shown in the conjugations of the aspectual pair **делать** / **сделать** 'to do' (1st Pers. Sing. (**делаю** / **сделаю**), 2nd Pers. Sing. (**делаешь** / **сделаешь**), 3rd Pers. Sing. (**делает** / **сделает**), 1st Pers. Pl. (**делаем** / **сделаем**), 2nd Pers. Pl. (**делаете** / **сделаете**) and 3rd Pers. Pl. (**делают** / **сделают**)). Imperfective future-tense verbs also have six forms and are formed by adding a verb infinitive to a conjugated form of **быть** 'to be' (**буду**, **будешь**, **будет**, **будем**, **будете**, **будут**). In the past tense, verbs, both imperfective and perfective, have four forms distinguished according to gender and number: masculine singular (**делал** / **сделал**), feminine singular (**делала** / **сделала**), neuter singular (**делало** / **сделало**) and plural (**делали** / **сделали**). In addition, all verbs have imperative (**делай(те)** / **сделай(те)**) and conditional forms (formed by adding the particle **бы** to the past-tense form of a verb: **делал бы** / **сделал бы**), and many aspectual pairs have four participle forms (present active (**делающий**), past active (**делавший** / **сделавший**), present passive (**делаемый**) and perfective passive with distinct long and short forms (**сделанный** / **сделан**) and two gerunds (imperfective (**делая**) and perfective (**сделав**)).

5 Statistical tagging and lemmatization

Because of the considerable amount of morphological variation in Russian, mapping forms to their lemmas (dictionary headwords) is not straightforward. In addition, the level of syncretism is relatively high: forms can usually have several grammatical interpretations depending on the context; the same is observed across part-of-speech (POS) categories – for example, **мой** is both a possessive pronoun (meaning 'my') and the imperative form of the verb **мыть** 'to wash'. Statistical tagging assigns the most probable tag to the next word given a sequence of n (usually $n = 2$) previous words (see chapter 5 in Jurafsky and Martin, 2008). Once the tag is known, the lemma can be derived using the list of forms with their

tags. The ambiguity in this mapping also depends on the set of tags used by the tagger. If a tagset can discriminate between the major syntactic classes (e.g. pronouns vs verbs), we can detect whether the form **мой** has the reading 'my' or 'wash' in a given context. However, a tagset distinguishing between only the basic parts of speech is not capable of lemmatizing word forms like **банки** or **физику** to the right lemma, because these forms have both masculine and feminine readings, which map to different lemmas, **банк** 'bank' vs **банка** 'jar'; **физик** 'physicist' vs **физика** 'physics'. A more extensive tagset distinguishing nouns by their gender can do this task (provided that the tagger assigns the right tag).

We have a reliable POS tagger and lemmatizer (Sharoff et al., 2008), which has been used to process I-RU. The corpus used for training the tagger was the disambiguated portion of the Russian National Corpus (Sharoff, 2005). The accuracy of tagging is about 95 per cent and the accuracy of lemmatization more than 98 per cent. However, we checked the I-RU-derived frequency list manually. Grammatical aspect is an area of Russian grammar that English-speaking students fail to assimilate fully.

The translations of a verb in the two aspects are usually quite similar, so lemmatization mapped the closely related aspectual pairs (e.g. **бросать** / **бросить** 'to throw') into one entry corresponding to the verb in the imperfective aspect. However, we avoided doing this for the perfective verbs produced by prefixation (**делать** / **сделать** 'to do') or having an irregular pattern (**говорить** / **сказать** 'to say'). Both verbs are listed in the dictionary in such cases. We have also unified many fine-grained distinctions made for uninflected forms, i.e. cases in which the difference in the syntactic function of a word has no overt morphological expression, e.g. **пусть**, 'let' as a conjunction and as a particle. Many native speakers fail to make such distinctions; the same applies to the language learners and statistical POS taggers. Finally, for this dictionary we also unified the adjectival nouns with their respective source adjectives; for example, **гласный** 'vowel' and **русский**, 'Russian'. This decision was partly determined by the similarities in their meaning, and partly again by the less reliable detection of this distinction.

6 Creating the dictionary

We started with a rough frequency list of the lemma-POS pairs in I-RU. For the purposes of compiling this dictionary we deleted from this initial list all the proper names (e.g. **Владимир** 'Vladimir' and **Газпром** 'Gazprom') with the exception of the most common geographical names, which are likely to benefit beginners (**Москва** 'Moscow'). Towards the end of the list we applied more filtering by removing trivial morphological transformations (e.g. **республиканский** 'republican', since it can be easily derived from **республика** 'republic') and words that are likely to be of little interest to the general language learners except those studying specialized domains (**дупло** 'tree hole').

The lemmas were ranked by their frequency (normalized as instances per million words). We also computed Juilland's D coefficient (Juilland et al., 1970), which represents the dispersion of frequency across the range of documents:

$$D(x) = 1 - \frac{\sigma(x)}{\mu(x)}$$

where $\sigma(x)$ is the standard deviation of the normalized frequency of word x over the documents in the corpus, while $\mu(x)$ is the overall average frequency of this word. The value ranges from 1 ($\sigma = 0$), i.e. a word is equally frequent in all documents, to 0, when a word is extremely frequent in a small number of documents. In this dictionary we multiply this value by 100 for typographic reasons.

A technical issue concerns the use of the letter **ё** (yo), which is normally written as **е** in standard Russian texts except those intended for children and foreign language learners. Given that the letter is not marked in the vast majority of Russian texts and is rare in our corpus, it is only marked in the headword, while we have not adapted the examples from the corpus. In most of the examples **ё** is written as **е**, but readers can work out where **ё** occurs from the headword. In addition to ranking the top 5,000 individual words, we included the 300 most common multiword constructions consisting of two or three words. Formulaic language is very important for language learners (Biber, 2009). Furthermore, many Russian constructions make sense only taken as a whole, e.g. **друг друга** ('each other', lit. 'friend (to, of) friend'). For this task we

started from an initial list of the most common two- and three-word expressions ranked by the log-likelihood score (Dunning, 1993) and then selected a pedagogically relevant list.

The examples in the dictionary entries were selected from the same corpus, from which the frequency lists were extracted. We aimed at selecting representative examples in which the headword is used with its most significant collocates as detected by the SketchEngine. A word normally has a number of contexts of use. In some cases, we selected more than one example per headword to illustrate very different contexts, but in this dictionary we did not have space to cover all words. In selecting the examples we balanced the need to illustrate the most common patterns of use vs the need to show the 'basic' sense of the word, from which more metaphorical senses can be derived (even if a metaphorical sense is itself more common than the literal sense). All the examples have been taken from the corpus. However, in many cases we have adapted the authentic examples to shorten them, to reduce the amount of unfamiliar words or to remove less common syntactic constructions.

Translation of the examples revealed many aspects specific to Russian language or culture. In a short isolated example it was often difficult to give justice to the connotations of a particular expression, while keeping the same structure as in the original Russian example. It is also useful to expose students to the differences between the syntactic structures expected in English and in Russian. Therefore, we tried to provide the most fluent translation, even at the expense of deviating from the precise wording of the Russian.

For example, for illustrating the use of **такой** as an intensifier, the Russian sentence '**Почему у вас такой усталый вид?**' (lit. 'why with you such a tired look?') was translated as *Why do you look so tired?* The noun **вид** in this example was also translated as the verb *look*. Another case in point is the translation of the Russian term **Великая Отечественная война** (lit. 'Great Patriotic War') as *World War II* in our examples. While technically the two terms are not fully equivalent, learners should benefit from the possibility of recognizing the connotations of the collocation (e.g. **ветеран Великой Отечественной** 'World War II veteran').

7 Using the dictionary

The dictionary includes the following lists:

Frequency list The frequency list contains the 5,000 most frequent lemmas with the following information:

- rank order of frequency
- normalized frequency (per million words)
- headword (lemma) with stress given for polysyllabic words
- part of speech indication (with gender information for nouns)
- illustrative example from the corpus with translations into English
- Juilland's D dispersion index.

For example, the entry

2565 дача Nf summer home, dacha

- Она провела лето на даче — She spent the summer at the dacha.
39.8; D 95

indicates that this word has the rank 2565 in the frequency list, it is a feminine noun (Nf), its frequency in the corpus is 39.8 ipm (instances per million words), while the D value is 95, making it a lexical item reasonably well spread across the texts we have in the corpus.

Alphabetical listing This lists the 5,000 words in alphabetical order with the following information included:

- rank in the frequency listing
- lemma with part of speech
- English translation.

Part-of-speech listing This lists the words in the frequency order separately for the main parts of speech (nouns, adjectives, verbs, adverbs) with the following information included:

- rank in the listing for this part of speech
- rank in the overall frequency listing
- lemma.

Multiword constructions This lists 300 multiword constructions (consisting of two or three words).