Meaning as Use: a communication-centered approach to lexical meanings

Serge Sharoff
Centre for Translation Studies,
University of Leeds,
Leeds, United Kingdom, LS2 9JT
s.sharoff@leeds.ac.uk

19th February 2004
Abstract

This is a working draft of the book. Please send your comments to the author.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>1</td>
</tr>
<tr>
<td>I The Theoretical Exposition</td>
<td>7</td>
</tr>
<tr>
<td>2 The relationship between words and concepts</td>
<td>9</td>
</tr>
<tr>
<td>2.1 The two paradigms</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Words as Pointers to Senses</td>
<td>12</td>
</tr>
<tr>
<td>2.2.1 Mismatches between concepts and uses</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2 Mismatches between concepts in translation</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Words as Resources for Communication</td>
<td>26</td>
</tr>
<tr>
<td>2.3.1 The basic ideas of the meaning-as-use model</td>
<td>26</td>
</tr>
<tr>
<td>2.3.2 The phenomenological structuralism of Edmund Husserl</td>
<td>29</td>
</tr>
<tr>
<td>2.3.3 The relationship between Husserl’s phenomenology and systemic-functional linguistics</td>
<td>37</td>
</tr>
<tr>
<td>2.4 Communication-centred description in action</td>
<td>38</td>
</tr>
<tr>
<td>2.4.1 Case Study: telling and translating a story</td>
<td>39</td>
</tr>
<tr>
<td>2.4.2 Case Study: terminology in software manuals</td>
<td>46</td>
</tr>
<tr>
<td>3 The Form-Meaning interface</td>
<td>51</td>
</tr>
<tr>
<td>3.1 Principles of the corpus-based study</td>
<td>51</td>
</tr>
<tr>
<td>3.2 What we can find in corpora</td>
<td>54</td>
</tr>
<tr>
<td>3.3 Corpora used in the study</td>
<td>61</td>
</tr>
<tr>
<td>3.3.1 The Russian Reference Corpus</td>
<td>63</td>
</tr>
<tr>
<td>3.3.2 The aligned corpus</td>
<td>63</td>
</tr>
<tr>
<td>3.4 The frequency list for modern Russian</td>
<td>65</td>
</tr>
<tr>
<td>3.4.1 Resolving the ambiguity of forms</td>
<td>66</td>
</tr>
<tr>
<td>3.4.2 Resolving problems with word uses</td>
<td>69</td>
</tr>
</tbody>
</table>
3.5 The conduit metaphors in translations .......................... 72

4 The lexical model ................................................. 79
   4.1 Introduction ................................................. 79
   4.2 Lexis in systemic linguistics ............................... 82
   4.3 The three views on lexical resources ....................... 84
       4.3.1 The view “from around” the lexicogrammar .......... 86
       4.3.2 The view “from above” the lexicogrammar .......... 87
       4.3.3 The view “from below” the lexicogrammar .......... 89
   4.4 Case Study: the description of “to burn” ..................... 93
       4.4.1 The lexical domain ..................................... 93
       4.4.2 burn and glow: the view from within the network ... 98
       4.4.3 burn: the view from below ............................ 102

II The Comparative Lexicogrammatical Cartography 105

5 Size adjectives .................................................... 107
   5.1 The methodology for description ............................ 107
   5.2 Compilation of the list of size adjectives .................. 109
   5.3 Detecting basic choices ...................................... 111
   5.4 Testing the basic choices .................................... 112
       5.4.1 English .................................................. 113
       5.4.2 Russian .................................................. 120
       5.4.3 German .................................................. 125
   5.5 Extending the network of choices ............................ 126
   5.6 Representing translation equivalence ......................... 136

6 Verbs of motion ................................................... 141
   6.1 Introduction .................................................. 141
   6.2 Basic options for using verbs of motion .................... 146
   6.3 The subcase of abstract motion .............................. 150
   6.4 The difference between logic- and communication-centred
descriptions ..................................................... 156
   6.5 The preservation of lexical features in translations ...... 158

7 The Core Lexicon of Emotions .................................... 163
   7.1 Introduction .................................................. 163
   7.2 The scope and methodology .................................. 167
   7.3 Collecting the list of emotions .............................. 171
## CONTENTS

7.3.1 Methodology ........................................ 171  
7.3.2 Results ............................................... 173  
7.4 The framework for uses of emotional words ........... 175  
7.5 Words referring to anger ............................... 185  
  7.5.1 The structure of the lexical field ............ 185  
  7.5.2 A case study of anger references in a story .... 191  
7.6 Conclusions ............................................. 194  

8 Conclusions ................................................ 195
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Formal and functional approaches to lexical semantics</td>
</tr>
<tr>
<td>2.2</td>
<td>The Lexical Matrix in WordNet</td>
</tr>
<tr>
<td>2.3</td>
<td>A more elaborate view on the act of communication</td>
</tr>
<tr>
<td>2.4</td>
<td>Huserl’s model of internal time consciousness</td>
</tr>
<tr>
<td>3.1</td>
<td>The diagram of the frequency distribution</td>
</tr>
<tr>
<td>4.1</td>
<td>Demo of a systemic network for mood in English</td>
</tr>
<tr>
<td>4.2</td>
<td>The two models of lexicon in SFL</td>
</tr>
<tr>
<td>4.3</td>
<td>The relationship between lexical choices and extralinguistic inquiries, size adjectives</td>
</tr>
<tr>
<td>4.4</td>
<td>Concordances for burn and glow in English</td>
</tr>
<tr>
<td>4.5</td>
<td>Concordances for goretj and svetijjja in Russian</td>
</tr>
<tr>
<td>4.6</td>
<td>The basic options for burn</td>
</tr>
<tr>
<td>4.7</td>
<td>Analysis of concordance lines in the Systemic Coder</td>
</tr>
<tr>
<td>5.1</td>
<td>The basic oppositions for uses of size adjectives</td>
</tr>
<tr>
<td>5.2</td>
<td>The concordance for large</td>
</tr>
<tr>
<td>5.3</td>
<td>The additional oppositions for uses of size adjectives</td>
</tr>
<tr>
<td>5.4</td>
<td>The classification of imaginary quantities</td>
</tr>
<tr>
<td>5.5</td>
<td>The new oppositions for dimensionality</td>
</tr>
<tr>
<td>5.6</td>
<td>A partial view of the directional-type chooser in KPML</td>
</tr>
<tr>
<td>5.7</td>
<td>A nominal group with a lexicalised size adjective</td>
</tr>
<tr>
<td>6.1</td>
<td>The basic options for verbs of motion</td>
</tr>
<tr>
<td>6.2</td>
<td>Basic options for abstract &quot;away-from&quot; motion in English, German and Russian</td>
</tr>
<tr>
<td>7.1</td>
<td>Basic options for expressing emotions</td>
</tr>
</tbody>
</table>
7.2 Options for expressing emotional states .................. 181
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>The composition of the Russian Reference Corpus</td>
</tr>
<tr>
<td>3.2</td>
<td>The frequency distribution for some Russian words</td>
</tr>
<tr>
<td>3.3</td>
<td>The frequency distribution of <em>admiral</em> in several texts</td>
</tr>
<tr>
<td>3.4</td>
<td>T-score values for the collocates of <em>get</em></td>
</tr>
<tr>
<td>4.1</td>
<td>The results for <em>burn</em> and <em>glow</em></td>
</tr>
<tr>
<td>5.1</td>
<td>The most frequent size adjectives in English, German and Russian</td>
</tr>
<tr>
<td>5.2</td>
<td>Annotations of English adjectives</td>
</tr>
<tr>
<td>5.3</td>
<td>Frequency values for small and little collocating with children</td>
</tr>
<tr>
<td>6.1</td>
<td>The most frequent verbs of “away-from” motion in English, German and Russian</td>
</tr>
<tr>
<td>6.2</td>
<td>The distribution of motion types in translations of “Alice in Wonderland”</td>
</tr>
<tr>
<td>7.1</td>
<td>Uses of emotion words with respect to parts of speech</td>
</tr>
<tr>
<td>7.2</td>
<td>The most frequent words in the field ‘peace-anger’</td>
</tr>
</tbody>
</table>
Chapter 1

Preface

— When I use a word, . . . it means just what I choose it to mean - neither more nor less.
— The question is, - said Alice, - whether you can make words mean so many different things.
— The question is, - said Humpty Dumpty, - which is to be master, that's all.

Lewis Carroll, “Through the Looking Glass”

As any scientific inquiry, the present book stems from fascination. My fascination in this study is caused by the fact that we can use words to mean so many different things and by the fact that nevertheless we can make sense of words used in so many different ways. WordNet (Miller, 1990) lists 11 senses of the word word and is pretty conservative in its estimate, because the Collins COBUILD English Dictionary (CCED, 1995), for instance, lists 37 senses of word. The number shown after each word in the following sentence from the epigraph corresponds to the number of its senses in WordNet:


A simple combinatorial test shows that this simple sentence consisting of only 13 words can have 7,464,441,600 semantic interpretations according to WordNet. This is again a very conservative estimate, as the words the, whether, you have no entries in WordNet, while CCED (1995) lists, for instance, 19 senses of the and two senses of each of whether and you. WordNet also does not list the sense of can as the modal verb (whereas CCED lists 12
separate senses of *can* only in this function), so the total number of possible interpretations of (1.1) can be much higher. However, in spite of the number of interpretations, the sentence is in no way ambiguous for the human reader. The same problem of the almost infinite potential and the limited actual set of choices applies to translations. The word *word* can be translated into Russian in many different ways. A typical English-Russian dictionary includes a long list of translations, including *slovo, izvestie, soobščenie, sluxi, tekst, zamechanie, zapis’, komanda, ssora*, etc, not counting translations of multi-word units, such as *God’s Word = Svyaščennoe Pisanie* or *four-letter words = rugatel’stva*. The central thesis of the present study is that we can describe the multitude of word uses in a succinct way, and make a multilingual computationally-tractable model that describes similarities and differences between word uses in several languages.

The book brings together various ideas that are important for understanding the nature of language-based communication. The list of relevant disciplines includes corpus linguistics, lexical semantics, systemic-functional linguistics, Husserl’s phenomenological philosophy, natural language processing and translation studies. At the first glance, this collection of theories looks like a collection of artefacts presented in the historical museum of a provincial town: the letter of city rights issued in 1250, a tapestry from 1750, and a machine tool developed by a local inventor in 1850. However, the collection in such a museum is organised around a single topic: it is the history of the town into which all the artefacts are interwoven. The same is true for the present book. The collection of theories is arranged around the single research question: *what is the nature of linguistic meanings?*

The book is devoted to the study of uses of words in relationship to meanings they convey. Some theories in lexical semantics assume that words have meanings as entries in the dictionary and refer to concepts that can be defined in a logical formalism. This view can be described as Platonicist semantics: all meanings of words are there in the ideal world of language, and dictionaries are (im)perfect means for capturing them. Another approach assumes that words serve our needs for communication. This view can be described as Humpty-Dumptean semantics: the speakers use words to achieve their communicative goals. In this view, words are described in terms of their contribution to the ongoing discourse. This also assumes that the study is based on examples of word uses taken in their context.

The second (communication-centred) approach has been widely used in sociolinguistic studies, however, it had relatively little currency in research on lexical semantics, especially in computational lexical semantics. The two approaches can be considered as complementary both in terms of their
research goals and resulted descriptions. As explained below they are related to traditions in human thinking existing for thousands of years, so that there is no estimation that one of them wins in the immediate future, much less as the result of the publication of the current book. However, the book advocates the communication-centred approach and indicates certain descriptive problems that we encounter, when we apply the logic-centred approach to studying uses of words in naturally occurring communication.

The book consists of two parts. The first part discusses principles of the communication-centred approach to linguistic meanings, presents a formal model to describe words in their context and compares the results against descriptions produced within the logic-centred approach. The second part is devoted to application of the theoretical exposition to studying lexical semantics in several domains, such as size, motion and emotions.

The major research question concerning the nature of linguistic meanings extends into several subtopics. The first extension: what are theoretical assumptions for our conception of meanings? This is the topic of Chapter 2, which studies the opposition of the two possible answers to the main question addressed by the book. In particular, the chapter discusses basic approaches within the logic-centred paradigm, detects cases, when logic-centred descriptions fail to account for uses of words, and outlines an approach for describing uses of words and meanings intended by uses. The communication-oriented approach advocated in the book is related to the phenomenological movement originated by Edmund Husserl. The chapter presents possible impacts of the Husserlean approach on modern studies in linguistics and cognitive science. The linguistic theory that provides the ground for the study presented in the book is systemic-functional linguistics as developed by Michael Halliday, so the chapter also compares the ideas of Husserl and Halliday.

The second possible extension of the major research question is: What are empirical foundations for the study of communication and how can computational tools help in the study? The topic of Chapter 3 is the relationship between forms and meanings. Since the communication-based approach is aimed at studying uses of words in their context, the study should be based on corpora, collections of texts that are representative for a language in question. However, any corpus study is based on forms of words, while the aim of a communication-centred study is to describe meanings conveyed by them. This requires interpretation of word forms from the viewpoint of their contribution to the discourse. The chapter starts with an introduction into principles for describing word uses on the basis of corpora, describes concordancing tools and corpora used throughout the book, and presents several
case studies concerning the interface between forms, which are stored in texts and available for corpus studies, and meanings, which are communicated by means of those texts, but not readily accessible in corpora.

The major contribution of the book is outlined in Chapter 4. The chapter proposes a computationally tractable mechanism for describing lexical semantics from the viewpoint of uses of words in naturally occurring communication. The mechanism distinguishes between (a) a set of lexicogrammatical resources offering the potential for communicating intentions of the speaker and for understanding utterances by the hearer, and (b) instantiation of the lexicogrammatical potential in the speech act. The theoretical exposition is followed by a case study (Section 4.4) that shows the communication-oriented solution to some of the problems identified in Chapter 2.

The third extension of the major question is: How far can we go in the description of meanings of from the communication-centred perspective? The second part of the book is devoted to three large-scale case studies, concerning uses of size adjectives, verbs of motion and words denoting emotions in three languages: English, German and Russian. The three descriptions differ in the level of details. The description of size adjectives in Chapter 5 (66 headwords in the three languages, covering 356 types of uses) is aimed to describe the complete domain with enough details concerning each single type of uses. The description of motion verbs in Chapter 6 (70 headwords, about 500 types of uses) concentrates on the most frequent types of uses verbs of “away from” motion. The description of emotion words in Chapter 7 (about 700 headwords) is aimed at covering the wide domain of words expressing emotions in very general terms with more attention paid to the domain of expressing anger and anxiety in the three languages (about 90 headwords).

The book refutes the dictionary metaphor: we use words and we understand words used by others in a way that is very different from the list of senses in a dictionary. However, the analyses offered in the book owe a great deal to existing dictionaries. After all, dictionaries reflect the traditions of uses of words, and set standards for speaking and writing. The three dictionaries actively used in the book are the Collins COBUILD English dictionary (CCED, 1995), the Longman Dictionary of Contemporary English (LDOCE, 1995), and the Oxford English Dictionary (OED, 1989). CCED (1995) is especially important, as it offers a comprehensive account of the English lexicon, which is functional and cognate to systemic linguistics: its descriptions relate words to functions of their uses. Also CCED is based on a representative corpus (the Bank of English), so its descriptions follow regular patterns of uses of English words (this is also the feature of
LDOCE). However, descriptions in a human-oriented dictionary lack formal mechanisms for dealing with meanings of lexical items in a computationally tractable way, i.e. such descriptions cannot be directly used in computational applications for language understanding and generation or machine translation. Also no large-scale dictionaries oriented towards functions of word uses and based on the corpus evidence exist for German and Russian.

The presentation of Russian words is a very technical issue, but itself it is a nice example of the form-meaning interface. Even though modern typesetting systems allow the direct use of Cyrillic characters, the shape of unusual characters, especially their italic versions given in examples, may distract the attention of readers not proficient with them. At the same time the number of schemes for transliterating of Cyrillic characters is beyond imagination. Wellisch (1978:256ff) lists a dozen of schemes only for transliterating Cyrillic letters used in Russian (not accounting for Cyrillic letters used in other languages). The relation of the topic to the form-meaning interface is that the form, i.e. a sequence of Latin letters in a transliterated string, tries to preserve the phonemic meaning of another form, a sequence of Cyrillic letters, but it may fail to do so faithfully, because of the difference between the number of letters in the Cyrillic and basic Latin alphabets and the ambiguity of some letter combinations. For instance, ts is frequently used for encoding ι, but it is legible as a sequence of original letters: исполь́зуется (ispol’zuets’ja), so this transliteration scheme introduces an ambiguity.

No transliteration scheme is perfect, but Russian examples in the book are transliterated using Latin characters without diacritics according to the standard transliteration scheme, System III (Shaw, 1967). This is a modification of the ISO/R9-1968 transliteration scheme geared towards linguistic analysis, because it considers the possibility of unambiguous reverse transliteration, but uses diacritics sparingly. It uses j for ʼă, ja for ʼa, y for ʼı, c for ʼu; for instance, vyjti encodes ʼvyjti.

The book could not appear without the help of various granting agencies. The conception of the book owns to the research support grant from the Open Society Institute, RSS 321/1999. The research continued in the course of the fellowship offered by the Alexander von Humboldt Foundation, Germany in 2001-2002 and the text was completed in the Centre for Translation Studies in Leeds.

Finally, I am glad to acknowledge the help of various persons with whom I discussed the ideas presented in the book. It is next to impossible to order and structure the list, so my thanks go in the alphabetic (i.e. almost random) order to John Bateman, Judy Delin, Dafydd Gibbon, Tony Hartley, Susan Hunston, Irina Kobozeva, Ekaterina Raxilina, Gert Rickheit, Lena Sokolova,
CHAPTER 1. PREFACE

Geoff Thompson, Constanze Vorweg and occasional discussions with many others.
Part I

The Theoretical Exposition
Chapter 2

The relationship between words and concepts

2.1 The two paradigms

In the preface I refer to the two approaches to the description of the meanings of words: the meaning is a concept or the meaning resides in the use. However, the distinction is not unique to lexical semantics. In different sciences that are aimed at studying the human mind we can often find two types of broad paradigms, which are complementary both in terms of their purposes and their results. The first paradigm is aimed at the logical analysis of structures that are necessary for cognitive operations. In the study of language, this paradigm is instantiated as formal linguistics, according to which language is considered as a set of rules for specifying grammatical structures, such as the construction of a transitive sentence with ‘verb + object’.

The second paradigm is aimed at the investigation of mind at work, when we make sense of objects in the environment, make sense out of messages issued by others or make senses by saying something to others, in short, when we interact with our biological or social environment. In contrast to the logic-centred paradigm, the paradigm of the second type can be called eco-logical. In the study of language, the ecological paradigm is instantiated as functional linguistics, according to which language is considered as a resource for communication, i.e. a resource for exchanging meanings by means of words. Matthiessen and Bateman (1991:54–55) refer to such attitude as rhetorical or ethnographic. The construction of a transitive sentence in it is described in terms of the Process and its Participants. In philosophy of
mind, the instantiations of the two paradigms are rationalist and empiricist
approaches respectively. In the case of lexical semantics, the first approach
assumes that a word has a meaning, which is defined as a sense in a dictio-
nary. The second approach assumes that the meaning of a word is a function
of its use in purposeful communication.

The first, logic-centred, paradigm assumes that meanings are concepts
that belong to an ontology, which represents real-world objects and their
properties. Lexical items may refer to one or several concepts and by virtue
of this reference they are endowed with a meaning, which also may depend
on truth conditions of a specific utterance. The ontology may be either sub-
jective (private for each human being) or objective (the slogan for the latter
case is “meanings are not in the head”). The spectrum of approaches within
this paradigm is wide: its proponents include Frege, Fodor, Putnam in phi-
losophy, Mel’ˇcek, Miller, Wierzbicka, in linguistics. Some theories from this
paradigm are in opposition to others, they all share the same ontological
commitment: the relationship between words and respective senses is pri-
mary with respect to communication, since the ontology, as a repository of
senses, exists and the mapping from words to senses is defined before any
possible act of communication.

In contrast, the second paradigm assumes the primacy of communi-
cation, because it assumes that human languages not informal prototypes for
knowledge representation languages, but tools that enable both the com-
munication of experience and interaction in society. In addition, the naive
ontology, which corresponds to the life-world of a human being, is defined in
relation to communication and is modified in the course of communication.
In this view, words are hints which are used by the speaker to achieve their
purposes. This view can be defined as the meaning-as-use position. It is
also shared by a wide community of philosophers of language and linguists,
for example, Husserl, Peirce, Wittgenstein in philosophy, Halliday (1978),

If the meaning of a word depends on its contribution to the ongoing ex-
change between the speaker and the hearer, it should be analysed in terms
of its occurrence in a specific utterance taken in its specific context as pro-
duced by a specific speaker communicating to a specific listener. However,
language is a resource shared by many parties involved in communication,
so the intersubjectivity of language restricts the potential of ways, in which
a word can be used. On the one hand, people do not act as semantic dic-
tator like Lewis Carroll’s Humpty Dumpty. If interlocutors aim to achieve
successful communication, they pack their message in words the recepient
can understand. On the other hand, the speaker is indeed the master of
2.1. THE TWO PARADIGMS

The formal model: meaning as concept

| mean | MEAN₁ | MEAN₂ | MEAN₃ |

The functional model: meaning as use

Figure 2.1: Formal and functional approaches to lexical semantics

how they mean things by means of words, what aspects of the situation are highlighted, how their evaluation of events is expressed, etc.

The difference between the two models is depicted diagrammatically in Figure 2.1. The logic-centred model approaches, when they refer to lexical semantics of individual words, use the dictionary metaphor: a word used in a text refers to a set of concepts that correspond to the set of its senses in a dictionary. See the diagram on the left side of Figure 2.1: *mean* can refer to MEAN₁, MEAN₂ or MEAN₃ and in the context of an utterance, such as (1.1), it means MEANₓ. However, a dictionary typically lists many more senses of a word, e.g. WordNet lists 15 senses of *mean*. The communication-centred model highlights the contribution of a word to the ongoing discourse. The speaker uses words to express his/her meaning intentions and the listener interprets the words being expressed to build his/her own understanding. Thus words, such as *mean* are used to encode speaker’s intentions in the context of the current speech act. Note also that the discussion shown on the right side of Figure 2.1 concerns the semantics of *it*. One of the interlocutors (the Mouse from Lewis Carroll’s “Alice in Wonderland”) assumes that that the meaning of *it* is fixed in much the same way as assumed by Frege or Wierzbicka.

There is also a terminological difference between the two approaches with respect to the topic of study. Logic-centred approaches prefer to use the term ‘sense’ to reinforce the dictionary metaphor. In functional approaches, the term ‘meaning’ is more appropriate, which follows the metaphor of communication: when a word is used, it carries a meaning intended by the speaker.
The two paradigms are contradictory: the first paradigm describes meanings abstracting from purposeful communication, while the second paradigm assumes that no such consideration is possible. At the same time, they are apparently complementary both in terms of their purposes and their results. Thus, they can mutually benefit from interaction between their proponents.

Sometimes, conceptually similar positions are advocated from the two opposite sides. However, the arguments used on each side are different. Marconi (1997:93–96) compared the arguments of Frege and Wittgenstein, who both rejected the possibility of a private language. Frege claimed that meanings are not private, because they are same and accessible for all speakers. Otherwise, the words for “The sum of squares drawn on its sides is equal to the square drawn on its hypotenuse” would not stand for the Pythagorean theorem but for my own theorem about right triangles, which could be different from your conception of triangles, squares and their sums. Thus, for Frege, meanings are public in the sense, in which tables and lemons are public in the material world.

Wittgenstein (1967) advocated the very same thesis but using another argument. For Wittgenstein meanings are not private, because they can only function in the society. To assert something privately is to assert nothing at all. So, meanings are public in the sense, in which traffic regulations or criminal laws are public.

The two subsections below offer a closer look at the two paradigms especially with respect to the conceptual system they offer for modeling uses of words in communication. The intention of the presentation is to show that even though the approaches are complementary, the communication-centred approach is better suited for analysis of uses of words taken in their context.

2.2 Words as Pointers to Senses

According to the logic-centred paradigm, the relationship between words and meanings is a one-to-many mapping between a lexical item and the set of its senses. Each concept in the list of senses is a separate item with necessary and sufficient conditions for considering a use as an instance of this concept. The difference between approaches in this paradigm concerns their understanding of how to represent concepts formally.

The simplest approach in logic assumes that each word in a language corresponds to a predicate denoting a concept, so that the meaning of chair is CHAIR and it can be used in logical formulas as such: SIT(x, y) & CHAIR(y). WordNet elaborates this approach using the idea of the lexical
2.2. WORDS AS POINTERS TO SENSES

matrix, in which senses are encoded by synsets, i.e. sets of synonymous lexical items (Miller, 1990). Figure 2.2 demonstrates the principle of the lexical matrix: all senses that exist in English are listed as $M_1$ \ldots $M_m$; all English words are listed as $F_1$ \ldots $F_n$. If there are two entries in the same column of the matrix (e.g. $E_{1,2}$ and $E_{2,2}$ in Figure 2.2), the word form ($F_2$) is polysemous; if there are two entries in the same row (e.g. $E_{1,1}$ and $E_{1,2}$), the two words ($F_1$ and $F_2$) are synonymous in this particular sense. For each lexical item, WordNet lists its synsets and often gives them a definition in informal English and some illustrative examples. For instance, a synset for *chair* in the sense of furniture contains one word:

(2.1) chair – (a seat for one person, with a support for the back)

A synset for *leave* in WordNet 1.6 has several words:

(2.2) leave, go forth, go away (go away from a place) “At what time does your train leave?; She didn’t leave until midnight”

However, Miller (1990) insists that definitions given in parentheses are used only for clarification of the nature of a synset, but a synset in WordNet is not defined, but signified by the set of its lexical items.

More delicate approaches to the semantic structure of senses assume that a sense has an internal structure. For instance, in formal semantics, the sense of the word *chair* can refer to its properties (Dowty, 1979:39):

(2.3) object(x) & physical(x) & \ldots & seat-for-one(x)

<table>
<thead>
<tr>
<th>Word Meanings</th>
<th>Word Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>$E_{1,1}$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$E_{1,2}$</td>
</tr>
<tr>
<td>$M_3$</td>
<td>$E_{2,2}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$M_m$</td>
<td>$E_{m,n}$</td>
</tr>
</tbody>
</table>

Figure 2.2: The Lexical Matrix in WordNet
Other approaches posit a more elaborate inventory of devices for formal representation of senses of lexical items. Most typically this is based on semantic networks consisting of a set of elementary nodes (atoms) and a set of relations for connecting nodes into larger units (molecules of meaning). The representation power of semantic networks is formally equivalent to predicate calculus, because both nodes and relations can be expressed in terms of predicates and logical connectors (Simmons, 1973). However, the mechanism of semantic networks offers a better level of abstraction, because it allows more explicit representation of the hierarchy of types, the relationship between types and instances, as well as between the predicate and its arguments. For instance, the concept ‘chair’ can inherit from ‘furniture’, which in its turn inherits from ‘physical objects’. The mechanism of semantic networks is used in many computational descriptions of lexical semantics, for instance, in the theory of conceptual graphs (Sowa, 1991). A longer review on the use of semantic networks to represent meanings is provided by (Lehman, 1992).

Linguistic counterparts of semantic networks use explications, which are similar to dictionary definitions, but are arranged on more formal grounds. Two famous approaches are the Natural Semantic Metalanguage, NSM, (Wierzbicka, 1980), (Goddard and Wierzbicka, 2002), and the Meaning-Text Theory, MTT, (Mel’čuk, 1988).

NSM describes meanings of words using a fixed set of semantic primes, like FEEL, BAD, I, YOU, SOMEONE, SOMETHING, which are claimed to be available in all languages. The number of primes varies in different versions of the theory from 13 (Wierzbicka, 1980) to over 60 (Goddard and Wierzbicka, 2002). However, the general principle remains the same: a word has an invariant meaning, which can be described in an explication consisting of semantic primes.

The later versions of the NSM theory also use a structured explication format that includes:

1. a mental formula, e.g. ‘X felt something because X thought something’;
2. a prototype scenario: ‘something good happened, I wanted this to happen’;
3. an evaluation: ‘when this person (I) thinks this, this persons feels something good’;
4. implications of the scenario for the person (X), which is different from
"I" in the prototype scenario: ‘X felt something like this, because X thought something like this’.

The explication above stands for X was pleased. The explication for X was delighted, for instance, differs from it only in the scenario: “I know now: something very good happened; I didn’t know that this would happen” and the evaluation: “feels something very good”. The problem with application of the description to words used in their context is that the description says relatively little about the actual behaviour of lexical items and about cases in which they are used. A good explication reveals something essential for concepts designated by words, but it does not embrace the whole range of their (non-metaphorical) uses. For instance, the following example:

(2.4) Tom was delighted at the sensation he was creating.

implies that Tom knew that he created the sensation and wanted to do it (in contrast to the scenario for delighted quoted above). Another case is the typical polite expression:

(2.5) Thanks for the invitation. I’d be delighted to come.

This implies neither that “something very good happened” (at most, this means that I’m pleased with the invitation) nor that “I didn’t know that this would happen”.

MTT provides a specific lexicographical device for representing senses of lexical items, the Explanatory Combinatorial Dictionary, ECD. An entry for each separate sense contains the definition zone, in which the sense is defined in terms of semantically simpler components, which are either primitive or have their own definitions (they should be ultimately reduced to primitives). Mel’čuk (1988:173) defines the following principle: “In the definition of a lexical unit L, each component must be necessary, and the set of all components must be sufficient, for the definition to identify L uniquely in all imaginable uses.” The definition is expressed in a natural language (English, French or Russian), but the ultimate goal for representing senses in ECD is to use a formal knowledge representation mechanism, like predicate calculus or semantic networks, though, they have never been actually used in ECD-style definitions, cf. in the two large-scale combinatorial dictionaries for Russian (Mel’čuk and Žolkovskij, 1984) and French (Mel’čuk et al., 1984).

The following subsections illustrate two types of problems with the ontological commitment: how to decide which sense is to be selected for a given
2.2.1 Mismatches between concepts and uses

According to the logic-centred paradigm, when a word is used in an utterance, it refers to one concept that is selected from the list of its senses. The mapping is based on the assumption that all possible uses of a lexical item should be expressed explicitly in its definition, which lists necessary and sufficient conditions for it referring to each and every concept. The drawback of the model is the possibility that the list of senses can be simultaneously too short and too long, once real uses are considered. It can be too short, because there are typically more possibilities for using a word than defined by a concept. This topic has been widely discussed in linguistic research: it is an immensely difficult task to construct a “waterproof” definition with a set of necessary and sufficient conditions for considering specific senses. A carefully constructed definition often fails when confronted with possible uses, cf. the discussion of bachelor in (Winograd, 1976) or examples with burn/goret’ and pregnant below. At the same time, if the list of senses gets more elaborate to accommodate for all possible uses, it starts to be too long, since some uses can be considered as ambiguous, i.e. referring to more than one sense in the list, while they are not ambiguous for either the speaker or the hearer.

The length of the list of senses of a word depends on the word and the lexicographer. Some words are more polysemous than others. Some lexicographers are ‘splitters’: they tend to analyse words in as many senses as possible to classify different examples of uses, while others are ‘lumpers’: they tend to provide more broad definitions covering a variety of uses. However, the lists are typically long. For example, the verb burn has 15 senses in WordNet, 13 senses in (CCED, 1995), and 20 senses in the Random House Webster’s Dictionary (the senses for the noun, phrasal verbs and idioms are not included in the count). In the ECD framework, (Apresjan, 1991) also analyses 16 senses of the Russian verb goret’, which is a rough translation equivalent of burn.

With respect to the list of senses, a natural question is: what are conditions for considering a separate item in it? An interesting example for this question is presented by the list of senses of burn from Random House Webster’s, including:

4. to give off light; glow brightly: The lights burned all night.
14. to use as fuel or as a source of light: to burn coal.
15. to become discolored by the effect of fire or heat.

The sense burn$_4$ does not discriminate between different ways for burning: an electric lamp, a candle or the sun burn differently, and this fact should be represented in the conceptual system in some way, so the list could be extended with 4a, 4b, 4c, etc. This might be required in a multilingual description, because it is plausible that another language may differentiate between natural sources of light, such as the sun or stars, and human-made artefacts, such as a candle or a lamp. This is the case, for instance, in Arabic, where the two processes are treated as conceptually different and are realised by different verbs:

(2.6) Al-ˇSamsu SaaTi’a / Al-qamaru SaaTig
sun-fem burn-fem / moon-masc burn-masc

‘The sun/moon is burning (natural object)’

(2.7) Al-ˇSam’atu tašta’il / Al-miSbaaHu yašta’il
  candle-fem burn-fem / lamp-masc burn-masc

‘The candle/lamp is burning (artefact)’

Then, there is the possibility of a mismatch between concepts defined as senses and word uses. For instance, the definition of burn$_4$ from Random House Webster’s allows only to glow brightly, while the following example contradicts it:

(2.8) a lamp was found burning dimly in the tomb

On the other hand all these meanings are described by different definitions, which do not explicitly relate process (14), light generated by the process (4), and its effect (15), while the basic meaning communicated by the author can only vaguely to these concepts. For instance, cars and buildings in the following two examples were giving off light (burn$_4$) and were used as fuel (burn$_1$), but this was not the main intention of the people who started this:

(2.9) When I arrived one of the vehicles was still burning.
(2.10) Protesters set cars on fire and burned a building.

If we try to create a formal definition of burn (hence, a definition which can be used in an NLP application), we are confronted with two choices: either each of the senses is a specific concept or some of them constitute a
single joined concept. In the first case, the separation looks arbitrary, since
the list can be easily extended, due to various real-world constraints (like
different ways of burning), situational dependency of classification, cf. the
famous example of bachelor (Winograd, 1976), restrictions coming from the
linguistic context, such as aspectual properties of a verb in Russian. Also
we need additional representational tools to join modes of usage of all these
senses. In the second case, forming a single concept for several senses gives
an unnecessary mix of senses which may be unrelated in some situations,
for example, if we create a single concept out of burn\textsubscript{4} and burn\textsubscript{14}, we will
not be able to describe cases, when coal is burnt in a closed furnace, so the
light cannot be perceived.

Representation of interaction between meanings is closely related to
metaphoric mechanisms, which are routinely used in production of new
meanings of polysemous words. Random House Webster’s also contains
the following (basic) sense of burn:
9. to feel strong emotion: to burn with desire.

which can be naturally treated as a metaphoric extension of burning, even
though it should be described by a separate concept.

This is an exercise in logical analysis of the structure of senses of burn
in attempt to spot potential problems, when a word is considered to be an
instance of a concept. However, if we turn our attention to real examples
of uses, the problem of distinguishing between senses happens to be even
more complex. Lewis Carroll’s “Alice in Wonderland”, for example, contains
several expressions with burn, including:

(2.11) Alice started to her feet, . . . and burning with curiosity, she ran
across the field. . .

(2.12) A red-hot poker will burn you if your hold it too long.

Can ‘burning with curiosity’ be considered as the same type of concept as
‘burning with desire’? Both imply a strong emotion, but since the structure
of the two emotions is radically different (the wish to get something vs.
the interest in knowing something), the ‘feeling’ of the emotion is different
as well. This means that we have at least two extra concepts for a single
unambiguous use of to burn.

The example (2.12) is covered Random House Webster’s even by two
senses:
6. to produce or feel sharp pain or a stinging sensation.
7. to be injured, damaged, scorched, or destroyed by fire, heat, or acid.
However, each of the two senses covers a number of distinct concepts (to be injured by fire, to be injured by acid, to be damaged by fire, destroyed by fire, etc), while on the other hand, the particular usage in example (2.12) intends both pain and injury caused by intense heat.

It seems evident that all senses of burn mentioned above are conceptually related, and there are good logical reasons why they contribute to the polysemous meaning of the verb to burn, but the relationship is not specified explicitly in the definitions in Random House Webster’s. Senses in WordNet also are designed as separate synsets. The ECD introduces the apparatus of semantic bridges in definitions, i.e. two related senses of a lexical item must share a non-trivial semantic component. However, some bridges are introduced artificially, i.e. a component is introduced for the sole purpose of not splitting the entry into two homonymous lexical items. For instance, Mel’čuk (1988) add the component “X is very full of Y” for PREGNANT.1 (X is a woman or female animal, Y is a fetus) in order to share it with PREGNANT.2 (information object X is characterised by, contains or entails much Y/many Ys, as if X were very full of Y). Firstly, the bridge component is not always relevant for PREGNANT.1. For example, the joke Mackenzie and Mel’čuk (1988) themselves use as their starting point implies that a woman claiming a seat is pregnant for only ten minutes. Secondly, the bridge component is not intended in many uses of PREGNANT.2. For example, the British National Corpus (BNC) lists (2.13)-(2.15):

(2.13) It is a dangerous moment, pregnant with hope teetering on the edge of despair.

(2.14) It’s a grey day, pregnant with rain.

(2.15) Oh which is the one that is pregnant with the terrible time?

Incidentally, none of the sentences (2.13)-(2.15) refers to information objects, thus violating necessary conditions in the definition of PREGNANT.2.

Let us not forget that the design of a dictionary entry as a set of separate senses is not universal. The concept and the format of modern monolingual dictionaries depend on the history of their development and on the function they serve in society. As Kilgarriff (1997a) points out, origins of the tradition of sense enumeration in a monolingual dictionary are related to the development of printed discourse, particularly new periodicals, and the growth of a literate population in England at the beginning of the eighteenth century. This brought about a re-evaluation of the nature of meanings: dictionaries appeared as a response to the demand for an authoritative source
which helps in understanding a use of a word or in checking/ensuring the correctness of its use. As the result of this intention, a dictionary entry is typically designed as a list of senses that denote objects or their properties. The mismatches between uses of words and senses discussed above make no harm to human users, because they interpret dictionary definitions according to their own needs, most typically to get advice on the possible meaning of an unknown word or a known word in a new context, or to receive advice on the use of a word they already know.\(^1\) However, the mismatches between concepts and uses of respective words create a significant problem for the meaning-as-concept model as a theory of lexical structure.

A step towards a more dynamic account of word meanings is offered by the Generative Lexicon (GL) theory by Pustejovsky (1995). Pustejovsky explicitly rejects the “sense enumeration” model of traditional dictionaries. According to GL, polysemous senses are represented by combining regular uses into an underspecified type, which is made more specific in the context. For instance, the word `book` may be interpreted either as a physical object or as a piece of information, because the function of a book is to be read by a person (this is represented by the Telic role) and it is produced by another person (this is represented by the Agentive role). So the type `book` can be defined as follows:

\[
\begin{align*}
\text{book}(x) \\
\text{formal} &= \text{phys.obj}(x) \\
\text{telic} &= \text{read}(P, y, x) \\
\text{agentive} &= \text{wrote}(T, w, x) \\
\text{constitutive} &= \text{bind}(x, z)
\end{align*}
\]

However, even though Pustejovsky’s theory is aimed at describing the dynamic account of sense production instead of sense enumeration (thus, it considers how a word is used in its context), it still belongs to the logical side of the continuum, since meanings of a word in GL are considered from the viewpoint of their logical structure, not from the viewpoint of their use in communication and possible intentions of the speaker. Indeed Pustejovsky’s proposal is similar to the notion of regular polysemy introduced within the MTT framework, cf. (Apresjan, 1974). At the same time, the particular set of senses of the noun `book` is not determined solely by the logical structure of the concept BOOK. For example, the sense `book_5` in WordNet is “a record

\(^1\)See also an empirical study on the practice of dictionary use by EFL students in (Nesi, 2000)
of commercial accounts”, which is a separate sense related to other senses of this word, but its GL definition is indistinguishable from the definition for *book*\(_1\) (printed work) cited in (2.16), because *book*\(_5\) is also a physical object, produced by writing and aimed at reading (though of a different sort), so (2.16) joins two apparently different concepts. Admittedly, *book*\(_5\) is always used in the plural and other dictionaries do add a restriction (pl.), but it is harder to express this fact in the more formal structure of WordNet, which is designed as a mapping between the list of words (defined as lemmas) and the list of senses (cf. Figure 2.2). In addition, *book*\(_8\) from WordNet (a book of tickets or stamps) is not produced by writing and is not aimed at reading, and even though it is constituted by pages bound together like *book*\(_2\), the notion of pages, the way they are bound and their function in *book*\(_8\) is different.

In short, the cognitive plausibility of the dictionary-based model of word usage can be questioned: are real uses of words by humans based on the distinction between senses? The set of senses of a lexical item in a monolingual dictionary is an informal prototype for the description of cognitive operations performed by humans. The design of WordNet is claimed to be inspired by “current psycholinguistic theories of human lexical memory” (Miller, 1990). However, how many senses of *burn* or *make* are available to a person to select from, when they encounter the word in a text? An average sentence has billions of possible interpretations, if we take into account all the dictionary senses of words (see example (1.1) from the preface). However, all of the interpretations but one are not perceived by most people as possible (with the exception of deliberately ambiguous sentences).

Also, it is unclear whether the metaphor of human lexical memory as a dictionary is plausible in cases, when words are felicitously used in a sense which is not covered in a dictionary, like WordNet, or when a non-ambiguous use corresponds simultaneously to several senses in a dictionary. For instance, when a word is used in a novel way, how is the reference to a non-existent sense established? There are different possible situations, in which this happens, for example, because of the need to express a new meaning or because of a mistake or a slip of the tongue or substandard linguistic habits of the speaker. Finally, the hearer with limited knowledge of language, e.g. a non-native speaker, is able to understand the intended meaning s/he never heard before. Communication is often robust with respect to deficient or nonstandard uses, such as the use of disinterested to refer to the lack of interest in a subject:

\[
(2.17) \text{He'd been so disinterested in her progress during six years of }
\]
22CHAPTER 2. THE RELATIONSHIP BETWEEN WORDS AND CONCEPTS

primary school that he’d only ever visited the school once.

Also, people engaged in real communication do not have a direct access to clear-cut instances of senses. They can only access language form, while meanings are not given, but should be inferred from the form. So, it is not clear how concepts, especially those that do not allow ostensive definitions like *justice* or *obscure*, can be acquired, because such concepts are rarely defined clearly in natural communication in a way similar to NSM or ECD definitions. It is also not incidental that WordNet lacks entries corresponding to more abstract grammatical concepts. For instance, such words as *the*, *whether*, *you* (see example (1.1) in the preface) are not listed in WordNet and not defined in the ECD, as it is much harder to define the concepts behind such words and they are much better described in terms of functions of their uses.

Similar problems are also relevant for natural language processing. One of the initial intentions of ECD was to provide a source for development of machine translation and language understanding systems, but computer applications that follow the MTT paradigm do not use formal definitions in the ECD style (Apresian et al., 2003). Also, when meaning-as-concept approaches describe real linguistic resources in terms of formal semantics, the set of senses is typically reduced, because there are no good computational mechanisms to represent the concepts and relationships between them corresponding to the 15 senses of *burn* in WordNet. So, the multiple senses of a word according to a dictionary are grouped into a few rough senses, typically, corresponding to homonyms or significantly different translation equivalents in the case of a multilingual application.

In addition, specific applications often require new senses, which are not covered by dictionaries and depend on the problem domain of an application or even the subdomain in a domain. For instance, in a description of X-ray films of the thorax:

(2.18) *On the left the pulmonary pattern is substantially intensified (arteries).*

the word *arteries* (in plural) refers to a visible deviation within the pulmonary pattern, not to arteries as organs of the body. Thus, this use of *arteries* can be considered as a specific sense in this domain, even though from the viewpoint of medical topology this is an organ of the human body.\(^2\)

\(^2\)The topic of domain-specific ontologies and their relationship to the lexicon is discussed in greater detail in (Kononenko and Sharoff, 1996).
2.2. WORDS AS POINTERS TO SENSES

2.2.2 Mismatches between concepts in translation

The reference to a particular sense from the set of senses should also be preserved, when a word is translated into another language. From the viewpoint of the meaning-as-concept paradigm this implies that the inventory of concepts is essentially shared across languages, which use different words for pointing to them. However, the task of meaning preservation in translation is not trivial.

In the worst case, there is a mismatch between concepts available in the two languages. For instance, the verb *to snap* is frequently used in English in the register of software manuals to mean an instantaneous event of a jump of an object being drawn into a new position:

(2.19) The arc snaps to the endpoint.

However, Russian lacks a verb, which can be considered as a good “translation equivalent” of *snap*. This leads to expressing the same meaning in a different way. Two options can be used in Russian translations: to refer to this event while it is happening (*prygnut’, jump*), or to refer to its result, when an object resides in a location. Software manuals prefer the second option, since the purpose for a snap action consists not in initiating a movement, but in setting a fixed point through which a line passes. Even in the latter case still there are two possibilities, involving two different types of presentation: *privjazyat’* (attach), as in example (2.20), or *proxodit’* (pass), as in example (2.21):

(2.20) duga privjazana k konechnoj tochke
arc attach-psv to ending point-dat
‘The arc is attached to the ending point.’

(2.21) duga proxodit cherez konechnuju tochku
arc pass-3sg through ending point-acc
‘The arc passes through the ending point.’

The concept underlying such uses of *to snap* is one and the same in the source text and in translation, however, different presentation devices should be used for referring to it in Russian.

Another problem with translation refers to the possibility that words in different languages offer different levels of granularity for referring to a concept. For instance, the Collins English-German dictionary distinguishes 19 senses for the translation of *leave* into German and lists 29 different
words which can be used as translation equivalents. Some differences between translation equivalents are related to the inherent ambiguity of *leave* in English: it does not specify the means of leaving, while this is often obligatory in German and Russian. For instance, *When he left Rome* should be disambiguated in German and Russian translations:

(2.22) \[\begin{align*}
\text{Als er von Rom wegging / wegfuhr / abflog} \\
\text{Kogda on iz Rima ushel / uhal / uletel} \\
\text{When he from Rome walked / drove / flew}
\end{align*}\]

The German translations of *to leave home* also depend on the fact, whether this means a permanent leave (in this case *weggehen* should be used) or a short-term leave (*wegfahren*). Thus, the need to translate a word into a target language multiplies the number of its senses.

Finally, even if a use is not ambiguous, it can refer to slightly different concepts in different languages. For instance, actions with buttons and choice lists as control elements of software applications are typically expressed in English by using the verb *choose* referring to a logical choice among available options:

(2.23) *Choose the Save button*

(2.24) *Choose File:Save*

However, in Russian, choosing a semi-real button in a dialog box is expressed as a physical action of pressing:

(2.25) \[\begin{align*}
\text{Nazhmite knopku Save} \\
\text{Press-imper button-acc Save} \\
\text{‘Press the Save button’}
\end{align*}\]

However, the action of choosing an item in a menu is expressed as a logical choice:

(2.26) \[\begin{align*}
\text{Vyberite punkt File:Save} \\
\text{Select-imper item-acc File:Save} \\
\text{‘Select the File:Save item’}
\end{align*}\]

*Click* is a more specific verb for the action referred to in (2.23), but it assumes that the action is performed using a mouse, while *choose* allows a possibility to use a keyboard for doing this. Even though, the verb *vybirat’*
2.2. WORDS AS POINTERS TO SENSES

in Russian is the typical translation equivalent of the verb *choose* in many contexts, it is not appropriate in (2.25), most probably because it does not imply an act of option selection together with its functional consequences, as in *choose*. Thus, the senses of the verbs *choose* and *vybirat'* in (2.23) and (2.25) should be represented by two different concepts. Such minor differences between senses of similar words in different languages lead to the separation of sets of their senses.

This problem also received attention of researchers from the logic-centred paradigm. In her analysis of human emotions Wierzbicka herself emphasises the cultural specificity of words denoting emotions. Wierzbicka (1998) describes differences between *anger* in English and *gnev* in Russian. The Oxford Russian dictionary (an authoritative and reliable source) lists *gnev* as the only translation equivalent of *anger*, however, the two words have different meanings. Unlike *anger*, *gnev* has the components of ‘strong emotion’, ‘justifiable reaction’ and ‘intention to act’, or in the NSM terms ‘this person feels very bad and wants to do something in response, because someone has done something bad against this person’.\(^3\)

The title of another paper using the NSM approach (Durst, 2001) is quite provocative: “Why Germans don’t feel ‘anger’ ”. The reason for his claim is that even though *anger* in English and *Ärger* in German are etymologically related and can be used as translation equivalents in some contexts, there is a fundamental difference between them, namely, that German expressions using *Ärger* typically occur in the context of a prepositional phrase using *in* (in), *mit* (with) or *über* (about) that expresses the cause of anger, for instance, *Ärger mit dem Chef, Erbe, Hauseigentümer, über die Nachbarschaft haben* (to have anger with the boss, will, landlord, about neighbours). So, according to Durst, *Ärger* refers to the emotional state directed to its cause, whereas *anger* refers to the state of affairs per se without necessarily mentioning its cause.

Also, the semantic field of anger in German is structured differently. In addition to *Ärger*, there are two other nouns *Wut* and *Zorn* with respective adjectives (*wütend* and *zornig*): *Wut* is used for referring to strong cases of anger, when one looses control over one’s behaviour, whereas *Zorn* emphasises the right to have the emotion, so it is frequently used in such collocations as *in gerechtem Zorn* (in righteous anger). Even though English has words that express similar emotions (*rage* and *wrath* respectively), their

\(^3\)Note that this assumes that the concepts of ‘to do something bad’ and ‘to feel bad’ are universal, while they are specific only with respect to culture, but also the situation and the immediate lexical context. See the analysis of *good* in example (2.33) and its Russian translations below.
semantics does not coincide with *Wut* and *Zorn*, which are everyday words frequently used in modern German for expressing respective types of anger. What is more, until recently, the two words were used as the two basic designations of this emotional state, while *Arger* as the generic designation of the emotional state, was less common. This observation is supported by the frequency of translation equivalents for anger in the corpus of European Parliament Proceedings. The most widely used equivalents for *anger* in it are *Zorn* (31%), *Wut* (21%), followed by *Arger* (11%), *Verärgerung* (10%), *Empörung* (7%).

However, if we accept the claim that *anger*≠*Arger*≠*gnev* and that they refer to culture-specific concepts, then no translation between English, German and Russian is possible, unless readers of translation know the exact differences between the concepts in respective cultures. In this view words *anger*, *Arger* and *gnev* refer to pieces of NSM code that are not identical. At the same time there is a range of translations of respective words based on the dynamical translation equivalence established in the contexts of their uses. This does not assume that there is an equivalence between concepts in the two languages, but that there are resources available in the target language that can create a meaning that is functionally adequate within the target language and culture. In such situation *anger* can be indeed fruitfully translated as *gnev* given a suitable context:4

(2.27) *she gave a little scream, half of fright and half of anger* . . .  
*Anja izdala legkij krik — ne to uzhasa, ne to gneva* . . .

(2.28) *swallowing down her anger as well as she could.* 
*s trudom sderzhivaja negodovanie* (lit. with efforts restraining her anger)

In example (2.27) *gnev* can be used as a perfect translation equivalent for *anger*, while in example (2.28) it is translated as *negodovanie*, which better fits the more or less fixed expression *s trudom sderzhivaja*.

2.3 Words as Resources for Communication

2.3.1 The basic ideas of the meaning-as-use model

The logic-centred approach makes no difference between the speaker and the hearer. Both have the same access to the Platonic world of language that

---

4The analysis of the semantic field of anger and anxiety in English, German and Russian is the topic of Section 7.5.
can be exemplified as the WordNet lexical matrix (2.2) and both know what a word means in any given context. The communication-centred approach assumes that the stance of the two parties involved in communication is radically different.

First, there is no immediate connection between the minds of the speaker and the hearer. The speaker has communicative intentions and uses lexicogrammatical resources to realise them in a sequence of perceptible signals, which can be acoustic signals, printed characters, non-verbal gestures or a combination of thereof. The hearer has an access only to the signals and interprets them according to his/her own intentions in the ongoing communication.

Second, there is always a difference between the interlocutors in their knowledge. After all, if the two parties have exactly the same knowledge, communication between them is of very limited value. From the logic-centred viewpoint communication is also restricted to the exchange of knowledge. However, the communication-centred approach adds the interpersonal dimension: words are used not only to exchange information, but also to maintain a community, see the death penalty discussion (2.37) in Section 2.4 below.

Thus, the success of an act of communication depends on the amount of shared knowledge, including:

- the degree to which the access to language resources is shared, for instance, the distinction between native vs. non-native speakers, for native speakers the difference can concern their social background, regional variations, etc;
- the degree to which the context of situation is shared, for instance, what the speaker does not know, while the hearer does;
- the difference in understanding of the set of possible communicative intentions, for instance, what are the reasons for the speaker to say something and for the hearer to listen;

The amount of knowledge shared between the two parties is the first condition of successful communication. The second important condition is related to mutual “empathy” on the both sides. On the side of the speaker, this includes the aspiration to communicate, as well as expectations on the (prospective) hearer and what s/he can understand. As Giles and Coupland (1991:60–61) discuss accommodation to the audience, which “can function to index and achieve solidarity with or dissociation from a conversational
partner, reciprocally and dynamically”. On the side of the hearer, “empathy” of successful communication includes the aspiration to interpret the content and the set of expectations about the speaker and the content s/he can deliver (for example, consider communication between students and professors).

These are sociolinguistic backgrounds of the communication-centred approach. Are they relevant for research in lexical semantics? If yes, how can we describe the meaning of an individual word in this model? The short answer is: when doing research in communication-centred lexical semantics we desire lexical resources for delivering communicative intentions in the case of production and for interpreting messages in the case of understanding (Figure 2.3). The relationship between words and concepts in this case is based not on the mapping from words to concepts, but on the realisation: words provide resources to realise communicative intentions of the speaker.

Words that realise communicative intentions function as hints, which are arranged by the speaker according to his/her expectations concerning perception of the message by the hearer in the current context (or implied context in the case of written communication). The signs are interpreted by the hearer according to his/her own access to the context. Here we can distinguish between two types of contexts: the largely extralinguistic context of situation and the immediate lexicogrammatical context in which signs are realised. Language resources provide a potential for communication. A subset from the potential is instantiated in the act of communication according to communicative intentions and the context of situation.

The fact that communication is typically successful, at least in cases studied in lexical semantics, leads to the impression that communication is transparent, i.e. concepts are directly given in communication and words are pointers to such concepts. However, the assumption of a direct mapping between forms and meanings is simplistic, because it is not possible to assume that there is a single and clearly identifiable concept behind each use, as we have seen in examples (2.8)-(2.15). A positive account of describing such phenomena from the communication-centred viewpoint will be given in the next sections. Now, we would like to refer to methodological flaws concerning the assumption of a direct mapping between words and concepts. This issue has been discussed in philosophy of meaning by several philosophers, in particular by Edmund Husserl.
2.3. WORDS AS RESOURCES FOR COMMUNICATION

2.3.2 The phenomenological structuralism of Edmund Husserl

When we talk about methodological foundations of the meaning-as-use approach, names that immediately spring to mind are those of John Austin (1962) and Ludwig Wittgenstein (1953, 1967). However, they focused more on aspects of language use as a social phenomenon, while Husserl paid more attention to the mechanics of the relationship between meanings and perceptible phenomena. So his views are more relevant for the study described in the book, even though, unlike Austin and Wittgenstein he devoted only passing attention to the topic of linguistic communication proper.

Edmund Husserl is acclaimed as the founder of phenomenology, a method for the description and analysis of consciousness on the basis of data of experience. The relevance of his ideas to modern debates in cognitive science and artificial intelligence has been noted by many researchers, see, for example, (Dreyfus, 1982, 1997), (Münch, 1993), (Baumgartner et al., 1996), (Sharoff, 2005). Husserl himself treated the project of phenomenology as a rigorous methodological basis for scientific research of the human mind, partly because of his background in mathematics (he did his PhD under
Weierstrass) and partly because of the general trend of the time (the end of the 19th–beginning of the 20th centuries). However, the phenomenological movement can offer more prospects for cognitive science and linguistics than a simple similarity of intentions.

The objective of Husserl’s phenomenology consists in the direct investigation of phenomena as they are grasped in consciousness. This attitude he expressed in the famous watchword Zu den Sachen selbst (To the things themselves). However, the slogan means neither the study of the objective world disregarding the human mind, nor simple introspection into our intuition. Husserl viewed this as is a suggestion to philosophers to turn their attention to data immediately accessible in consciousness, thus a phenomenologist in his view studies not things, but their meanings as they are experienced.

The development of Husserl’s views on the nature of meanings and language started with a theory presented in the second volume of ‘Logical Investigations’ (LI) from 1901, in which he analysed the structure of referring acts in the way similar to Frege’s Identification-Description model. However, from the very start, Husserl’s theory points out that linguistic meanings arise from acts of communication, in which they appear. The discussion held between Frege and Husserl at the end of the 19th century is yet another instantiation of the opposition between the two paradigms. For Frege, the Sense of a referring expression is independent from the act of communication and the parties involved in it. In Frege’s view, an expression simply has a meaning. In Husserl’s view, the meaning of an expression depends on “the sense-giving act character” (Husserl, 1952:303). In this respect, Husserl’s theory is act-based, while Frege’s theory is essence-based. For a detailed comparison of Frege’s and Husserl’s views, see (Smith, 1994) and other articles in the same volume (Haaparanta, 1994).

After LI Husserl developed a constitutive theory of meanings, which is based on intentionality, the directedness of mental states towards objects and states of affairs in the world. Even though such objects or states of affairs might be non-existent in the real world, they are always constituted in mental acts as meanings. This theory is expressed in the two books of “Ideas pertaining to pure phenomenology”: Ideas I, from 1913, and Ideas II, prepared by 1921, but not published in full until 1950s. However, in these studies, Husserl applied constitutive phenomenology to describing perception and the life world and made no clear indication how to apply it to linguistic signs and acts of communication. At the end of his life, he returned to the topic of creation, communication and reproduction of meanings in the society, in particular, in “Origins of geometry” (OG), from 1936. In the
view, expressed in OG, language transforms perceptual meanings into stable ideas, frees these ideas from perceptual contexts, establishes systematic oppositions between them and leads to communalisation of experience.

There are several recent treatments that elucidate Husserl’s approach to linguistic meanings. Welton (1983) analyses the development of Husserl’s approach to meanings diachronically and presents a reconstruction of Husserl’s ideas that relates phenomenology of perception to language as the intersubjective medium for communication. Smith (1990) gives another reconstruction of Husserl’s theory of linguistic meanings by considering it within the history of approaches to speech acts from Aristotle to John Searle. However, in what follows below I would like to extend these interpretations by proposing my own analysis that directly applies phenomenology to the study in lexical semantics.

There are two notions from Husserlian phenomenology that can be fruitfully imported into the research context of modern linguistics:

- intentionality as the correlation of noemata (phenomena given in experience) and noesis (a mode in which these phenomena are being experienced);
- the model of internal time consciousness (ITC), which assumes that the structure of all mental phenomena is based on their unfolding in time (Husserl, 1969).

For Husserl, all human perception and action is characterised by the directedness of consciousness towards objects and states of affairs in the world. Maurice Merleau-Ponty later used the term ‘the intentional arc’ as a feedback loop linking the subject and the object. All experienced phenomena exhibit a dynamic correlation between three parts of the intentional arc: physically perceptible features of an object being experienced [Gegenstand], the content of experience [Noema] and the mode the phenomenon is given in [Noesis].

In this model a human actively tries to make sense of physically perceptible features of objects given in perception. The interpretation does not go from scratch: noesis provides the context for constitution of noemata for perceived objects, but is itself constituted by them. At the same time every object in the intentional arc is grasped not as an abstract concept, but as a meaning (noema), which is relevant for the current noetic configuration. This means that we are active in making sense of objects we experience in the environment in a way that is relevant to current goals. This makes
Husserl’s noemata similar to affordances according to (Gibson, 1966): objects afford certain actions with them, for instance, a chair typically affords sitting and is not interpreted as a weapon, but it also affords to be used in this function in a case of danger.

Husserl’s noemata can be considered as meanings of linguistic expressions, and noesis as their context (McIntyre and Smith, 1982), i.e. we actively try to make sense of words we hear or read in a way that is relevant for our aims and the current context constituted by previous acts of interpretation. However, the intentionality of linguistic communication differs from the intentionality of perception in one important respect: the relation of words to real-world objects or notions is mediated by a public semiotic system of language, which is based on a system of oppositions and similes between signs, while perceptual meanings we live by (noemata) are based on their direct influence on our life. For example, the system corresponding to the word *red* defines possibilities for oppositions from things of other types and similarities to some related types: *red tomato, red apple, red hair, red wine, red light* are different instances of the same word *red*, as they are used in meaning-endowing or meaning-production acts, while perceptual noemata, i.e. meanings which are communicated by these acts, refer to different shades of this colour.

Thus, the semiotic system of human language adds an intermediate level: communication of meanings by physically perceptible signs is mediated by the lexicogrammar, which is a natural extension of the system of meanings that are directly experienced and communicated. At the same time it plays the crucial role in the development of that system.

The noema-noesis correlation is combined in Husserl’s view with another correlation, which links just-passed moments with the currently perceived now moment and anticipated future moments. Husserl’s model of internal time consciousness (Husserl, 1969) replaces the traditional single-dimension description of objective time by the two-dimensional schema which joins a sequence of objective flow of events with the way this sequence is grasped by consciousness (Figure 2.4). Husserl’s ITC model is based on a tripartite structure consisting of protention (t-future), the now-point, and retention (t-past). Retention is the chain formed of the past now-points, the reflections or adumbrations (Abschattungen) of previous phenomena kept in consciousness; in Husserl’s language, adumbrations continue to be “appresent”, present in the mode of retention. Protention is an anticipation of the future, the various expectations constituting and conditioning “that which is coming”.

In this view at the moment t-now consciousness grasps not only the
past, as a retention of \( t\text{-past1} \) and \( t\text{-past2} \), but also \( t\text{-future} \) is constituted at the moment \( t\text{-now} \) in the mode of expectation. The vertical line of the protention-retention chain in each moment (e.g. \( t\text{-now} \) or \( t\text{-past2} \)) modifies previous retentions, absorbs the just-passed now-point into the retention chain, and constitutes new protentions.

The integral model of noema-noesis correlation and ITC can help in the description of how the structures of the life-world are constituted through a sequence of meaning-endowing acts. At the real time moment \( t\text{-past2} \), the noema for \( t\text{-now} \) is constituted as an expectation on the basis the noemata of \( t\text{-past2} \) (which is “now” at that moment) and the retention chain. This constitution also conforms to possibilities available in the current noetic configuration. This noema, which is constituted in the mode of protention, is being realised (becomes concrete), so that it still keep its correlation to noesis, but probably alters its content. The content of a noema of a just-passed protention is altered, for example, when something unexpected has happened, thus violating previous acts of noematic constitution: the street we followed turned up to be a deadend. The same case is true for an expected but not happened event.

Noemata which have been constituted earlier in moments \( t\text{-past1} \) and
t-past2 are retained as a reduced content, which still keeps the features that are most salient from the standpoint of the current configuration of noesis, but may have lost some concreteness. In Husserl’s metaphor, they are visible “as through the layer of transparent water”. For example, when I listen to a lecture, in the midst of it I do not retain phonemes uttered in a previous sentence and the exact words expressed in its beginning, but I still keep in memory phonemes of previous words, the words of that previous sentence and meanings expressed at the beginning of that lecture.

Finally, noemata from the retention chain become hidden in this “water”, they sink into the long-term memory, they are no longer retractable in the short-term memory mode. However, they are stored in the long-term memory as a “packed” content of the sedimented intentional nucleus, which serves the basis for constitution of new expectations in presence of an associatively similar noetic configuration, which calls for particular features of sedimented nuclei (though only ready-to-hand aspects of a noematic nucleus are developed due to the condition of noema-noesis correlation). These new protentions are realised into the new actual t-now for a further sedimentation into a new noematic nucleus. This nucleus is enriched by new experiences and its new content is available for a further actualisation.

The ITC model is important for understanding of the way linguistic meanings are grasped in the flow of discourse, because the perception of text as a sequence of utterances has a temporal extension with elements that are incorporated into understood meaning structures of the text and expectations about various characteristics of the awaited speech, for example, if we have read to the following point:

(2.29) When you load a template or add-in, it remains …

there is a bunch of expectations about future elements to be found in the sentence, including structural constraints (they are typical for chart parsers) and semantic constraints, for example, a presupposition that the action expressed by the subordinate clause completed successfully, and the state resulted by its completion continues for some. The next of the clause may also have restrictions for its duration.

All this forms the presupposition potential that contextualises the rest of the clause and provides a basis for interpretation of next physically perceptible signs. Contextualisation implies two types of contexts, which are

---

5This mostly concerns understanding of the spoken discourse, because the written discourse is subjected to revisions upon writing and allows not strictly linear reading. However, the spoken discourse is the original mode of language use.
2.3. WORDS AS RESOURCES FOR COMMUNICATION

determined by language resources: the compositional context, which is set up by ways for composition of the tree of constituents, and the linear context, which is set up by ways for ordering of elements in development of the flow of discourse.

A result of meaning fulfilment of the presupposition potential may produce a meaning, which is identical to the protention (in the case of a completely predictable speech), partly satisfy its constraints or present a surprise. The realisation of this potential also does not involve infallible decoding with a direct access to intended meanings, because the sequence of physically perceptible signs is interpreted according to this potential. A professional software developer will understand (2.29) in a way different from a casual computer user.

The clause in example (2.29) finishes with:

(2.30) When you load a template or add-in, it remains loaded for the current Word session.

which closely matches the expectations.

The vast majority of signs are not introduced in a formal way using definitions. There is also no way of direct imprinting of new concepts that are meant by such signs into human minds. Thus, the only possibility for accessing new concepts involves their constitution in the course of one’s participation in meaning-maintaining practices of the society. Since a new sign (like add-in in the example above) is always contextualised within the ITC perception, it is endowed with a meaning, which may be not exactly the meaning intended by the author, but is relevant for meanings constituted for a sequence of adjacent signs. The event of meaning fulfillment for this sign leads to sedimentation of constituted meanings, which now contribute to the interpretation potential for future uses of this sign.\(^6\) So, the concept for add-in is gradually created, when one wields with such usages.

Even if the two ideas (the intentional arc and the ITC model) seem to be remote from research in modern linguistics, they are directly relevant to it. What is more they are relevant to the study presented in the book and they will continuously reappear throughout the book. First, the ITC model is immediately relevant to the idiom principle (Sinclair, 1991) or pattern grammar (Hunston and Francis, 2000), according to which the immediate lexical environment predicts the lexical item and its meaning.

\(^6\)Of course, this does not involve a conscious memorisation of all occurrences of a lexical item in the course of one’s life, but its recurrent usage in respective context is important for acquiring its meaning.
Second, in his description of clause complexes, Halliday also proposed a mechanism of “ball-chain” notation, which is necessary to capture “a continuous flow of [spoken] discourse that is coherent without being constructional” (Halliday, 1985:201–202). The mechanism, which is essentially similar to the ITC model, is designed to represent the ‘choreography’ of spoken language, which tends to develop the text in an evolving process, as opposed to the ‘crystalline’ tendency of written language to form ‘synoptic-type product’ structures of constituents and dependencies.

Finally, the two ideas are relevant for the very possibility of a semantic study based on the corpus evidence. When we use a corpus, the study is always based on physically perceptible features, namely, on forms of lexical items, as they are given in texts. At the same time, the goal of a study is to describe semantic phenomena underlying some (or all) types of uses of lexical items under consideration. Let us assume that we want to study uses of size adjectives. In this case we start with selecting adjectives for our study. On the one hand, the fact that an adjective, when it is used in a text, denotes a size of an object is a fact related to its meaning in this text. On the other hand, meanings are not readily available in a corpus, from which we can select, for example, expressions with adjectives referring to size. Moreover, there is no exhaustive list of communicative intentions that are realized by size adjectives. For instance, short is obviously a size adjective, even including cases when it is used for referring to time, as in short stay. What is more, the latter use is systematically related to size, so it should be considered in the treatment of size adjectives. However, this means that the complete description of size adjectives could cover adjectives that cannot refer to size proper, but only to temporal properties, e.g. brief. Thus, brief should be also included in the description. Incidentally, several types of uses of other words referring to temporal properties can also be included, for instance, young and old, because they are related to uses of size adjectives: his little brother, two small children.

When such words as fine are considered, the choice is less obvious. Even though, it is not considered as a size adjective, it is frequently used for this purpose, for example, CCED (1995) definitions “Something that is fine is very delicate, narrow or small” and “A fine detail or distinction is very delicate, small, or exact”. Such uses are relatively frequent: distinctions, needle, threads, tuning belong to the most significant collocates of fine according to the BNC. Moreover, when uses of translation equivalents of fine, are studied in German and Russian, they also often refer to size properties, e.g. fine distinction is translated into German and Russian as feiner Unterschied, tozke razlichie. This provides a reason for including fine in the
2.3. WORDS AS RESOURCES FOR COMMUNICATION

list of size adjectives to be studied, but at the same time, if the description concentrates on size adjectives only, it has to leave out other uses of *fine*, like “You use *fine* to describe something that you admire and think is very good” (CCED, 1995). The implications of the relationship between forms and meanings will be described in greater detail in Chapter 3.

### 2.3.3 The relationship between Husserl’s phenomenology and systemic-functional linguistics

Since the linguistic analysis offered in the book is based on Systemic-Functional Linguistics, SFL (Halliday and Matthiessen, 1999), it interesting to compare the ideas of SFL to Husserl’s phenomenology. The most of SFL writings themselves contain no clear reference to notions of Husserl’s phenomenology with a few notable exceptions. Halliday and Matthiessen (1999) occasionally refer to works by Thomas Luckmann, a member of the phenomenological movement, who was involved in phenomenological psychology and sociology. Also, Coppock (1997) compares Peircean and Husserlian phenomenologies and treats Halliday’s approach to social semiotic as essentially phenomenological, especially, with respect to the description of meaning maintaining practices in the society.

However, the relation between phenomenology and systemic-functional linguistics is both genetic and conceptual. A genetic link is provided by Roman Jakobson, who explicitly advocated the ideas of Husserl’s philosophy of language, see, for instance his study of the compositionality of meanings (Jakobson, 1971) and the analysis of Jakobson’s linguistic philosophy in (Holenstein, 1976). At the same time Jakobson was a leading functionalist, one of the founding members of the Prague Linguistic Circle, so that some of his ideas are also used in SFL, for example, see references to Jakobson in (Halliday and Hasan, 1985).

Conceptual links can be classified into three groups. First, SFL exhibits an “ethnographic” attitude, which implies that any work on grammar is based on naturally occurring texts taken in their context. This attitude reminds of Husserl’s watchword “To the things themselves”, which calls for the direct investigation of concretely experienced phenomena.

Second, both approaches assume centrality of meaning in the study of language and human activities, when objects and signs are analysed not only from the viewpoint of their physical appearance, but in relationship to their contribution to the constitution of meanings. Halliday’s analysis of grammatical meanings corresponds to the Husserl’s description of meaning-endowing acts of passive synthesis in which a noema is correlated with noesis. SFL
CHAPTER 2. THE RELATIONSHIP BETWEEN WORDS AND CONCEPTS

treats language as a resource for making meanings. This resource is organised in terms of three sets of functions (metafunctions), concerned with the representation of experience (the ideational metafunction), communication between the speaker and the hearer (interpersonal) and presentation of information as a text in context (textual).

The ideational and interpersonal metafunctions for Halliday are manifestations of the two very general purposes which underlie all uses of language: to understand the environment and to act on the others, while the textual metafunction is conceived as a linguistic realisation that breathes relevance into them (Halliday, 1985:xiii). Husserl, as early as in LI, also distinguishes an opposition between representational powers of linguistic acts and uses of language for asking questions, issuing commands, expressing admonitions, requests, etc. Husserl draws the second opposition between an actually experienced feeling, command or question, and the act involved in asserting a corresponding sentence, see (Smith, 1990).

Thirdly, both Husserl and Halliday exhibit an anti-hypostatic approach: meanings do not belong to an ontology which is independent from personal and social existence, but they are also not private for psychic subjects. Rather, such ontologies are produced and maintained in meaning-making practices of a community, hence they are intersubjective. In Ideas II, Husserl is preoccupied with processes involved in constitution of the lifeworld with the implicit impact on the acquisition of language. Halliday also assumes the primacy of meaning constitution with respect to all acts of cognition and describes language as the medium through which we make sense of our experiences and interactions with the world and society. One particularly important case of sedimentation of intersubjective ontologies is acquisition of the lexicogrammatical system of language, which is acquired through constitution and sedimentation of meanings which are brought in purposive communication activities. This links Husserl’s ITC model with the systemic-functional model of language acquisition according to (Halliday, 1978; Halliday and Matthiessen, 1999).

2.4 Communication-centred description in action

The following subsections present two case studies that elucidate the basic principles of the proposed approach to communication-centred description of lexical items. The first study describes the relationship between words and functions underlying their use in a story and its translations, while the second study considers the possibility to analyse terminology (for instance,
software terms) from the communication-centred viewpoint.

2.4.1 Case Study: telling and translating a story

One example of purposeful communication is a story, which is told in one language and may be retold in another one. We can consider three perspectives here. In the process of telling a story, the author uses the system of language as a resource for enabling the reader to create a mental image of events being described. Even in the case of a memoir or a story describing a real event, and even if its author tries to be faithful to facts, a story does not contain a complete representation of what happened in reality: any story delivers only a tiny portion of actions and events, their actors and objects and their properties, which are focused by means of linguistic expressions. It is the task of the author to choose what information is relevant to be conveyed and what is not. For example, the statement like She left yesterday does not specify the exact time, when she left, whether she left with luggage or not, whether she left by car or by plane, etc. However, a language may require to focus on a particular aspect. For instance, German and Russian require that the way of leaving is expressed explicitly, see example (2.22).

At the same time, a story is not confined to be just a representation. The very fact that a story is communicated means that the author tries to influence the hearer in some way: s/he presents a sequence of arguments and/or interpretations, expresses his/her own attitude or seeks an emotional response from the reader, asks rhetorical questions, etc. Finally, a story as a linear text has its own structural properties: what actors or objects are introduced and when, how the coherence of references to them is kept throughout the text, how the discourse is developed. In terms of systemic-functional linguistics (SFL), the three perspectives correspond to three groups of meanings, which are technically called metafunctions. Namely, the ideational metafunction corresponds to resources for representing the world; the interpersonal metafunction corresponds to resources for enabling the exchange between the author and the reader; the textual metafunction is conceived as a linguistic realisation of the first two metafunctions in the flow of discourse Halliday (1985:xiii). Halliday supposes that contributions of any utterance and of any constituent of an utterance, including separate words, into the discourse can be analysed in terms of those metafunctions.

Let’s start with the analysis of individual words. Logical analysis is mostly confined to nouns that categorize natural-kind objects. One particularly influential example is Putman’s analysis of the meaning of such words as water, elm or beech (Putnam, 1975). Putnam argues for the division of
linguistic labour: as a competent speaker of English, he knows the meaning of the word *water*, but leaves for a competent chemist the decision, whether a particular liquid conforms to the formula \( \text{H}_2\text{O} \) or not. The everyday usage should trust the decision of an expert. Prior to 1750 (this is roughly the time, when the water composition was detected) no such expert existed. Still the word *water* was used by English speakers, but the meaning was based on less definite identification procedures. According to proponents of logical analysis, the division of linguistic labour can be extended: there may be (1) an expert biologist that knows everything about a certain butterfly, say *Aulularia Clemensi*, but fails to recognize it in the Amazonian forest, and (2) an aboriginal that easily detects the butterfly type from its appearance and discriminates it against other types of butterflies but knows next to nothing about its biology (Marconi, 1997:58ff).

However, such natural-kind words as specific names of butterflies or even tree names take the minimal proportion in naturally occurring speech. The most delicate lexical choice for referring to *Aulularia Clemensi* is possible only in a very specific context of entomological studies, but even then, in a text that follows the proper identification, *Aulularia Clemensi* may be well referred in different ways according to textual metafunction resources available in a particular language and register. I was unable to find a text that refers to this butterfly type. It might be an invented example, such as many examples in the logic-centred paradigm. Let’s study a real example from an article in Encyclopedia Britannica that describes one of the butterfly families as (references to members of the family are underlined):

(2.31) *morpho* : (subfamily *Morphinae*), *any of a group of New World tropical insects* (genus *Morpho*) belonging to the brush-footed *butterfly family Nymphalidae* (order Lepidoptera). *(They are sometimes classified as a separate family, Morphidae.) Microscopic ridges on the butterflies’ wing scales break up and reflect light, producing the iridescent blue of some male *Morpho* species and the dull browns (brightened by contrasting eyespots on the wing undersurfaces) of male *Caligo* species. The generally duller coloured *females* have broader, less graceful wings than the *males*.

The list of nouns referring to this butterfly type includes, in addition to the explicit type name, such nouns as *butterflies, insects, species, they, females, and males* (i.e. female/male *Morpho butterflies*), etc. The reference is kept within the resources of the ideational and textual metafunctions. The ideational metafunction provides such classifications as ‘male’ vs. ‘female’,
or ‘butterfly’ is a subclass of ‘insect’. The textual metafunction provides possibilities of pronominal references or substitution of generic classes instead of more specific ones, if they are not ambiguous in the context. The description also employs the interpersonal stance, since it assumes what is known for an average reader of Britannica and what is not: the reader is expected to know that butterflies have wings, which widely vary in their colour and provide the basis for identification of butterflies. Butterflies are the subject of the article, so they are the grammatical subject of all sentences (except the description of microscopic ridges). In terms of the ideational metafunction, this fragment presents only a partial description of features of the Morpho subfamily, so it helps the reader in getting more knowledge about these butterflies. However, the image resulted from the information provided in the article is partial. It is insufficient for identification and may be extended by further reading.

Variation in naming of biological species may be caused by the complexity of the reference to them. Water is a much simpler type of objects and substances, and we have much more competence in referring to water than to *Aulularia Clemensi* butterflies. Still the meaning of the word *water* is quite complex and is not restricted only to the ability of establishing the equivalence between the substance that consists of H$_2$O and the word *water*, since the usage of the word *water* is defined not only or even primarily by the chemical composition of the liquid, but by the purpose, with which this word is used.

Winograd and Flores (1986:55ff) offer the classical example of the dialogue “Is there water in eggplants?”: if at a hot day somebody asks whether there is water in the refrigerator, then it is evident that s/he is thinking of water as a means to quench thirst. If there is a soft drink in the refrigerator, we can answer in the affirmative. Answers about water in the chemical content of eggplants or water condensed on the shelves might be relevant to an analytically-oriented logician but not to a real person in such a situation. The second answer, however, might be relevant to a technician who repairs refrigerators.

If the end of using language is to act on others, then the result of bringing a soft drink is sufficient for using the word *water* in this case. The discussion of water types at Earth and Twin-Earth in Putman’s thought experiment is taken out of context of purposeful expressions, by which we typically refer to natural kind objects. The lexical choice in a natural discourse is aimed not only at the proper identification of an object, but at purposeful interaction with the listener/reader to achieve communicative goals of the speaker/author. Thus, conditions of Putman’s thought experiment (un-
CHAPTER 2. THE RELATIONSHIP BETWEEN WORDS AND CONCEPTS

derstanding of the word water rests on whether it is a liquid of a specific chemical composition) miss the conditions under which we judge communication as successful and grant the opposite party with understanding of our intentions.

The situation with naming of natural-king objects is complicated further, when an object is described not by its class alone but by a nominal group with a complex structure. Halliday (1985:159ff) considers the following functional structure of a nominal group in English:

\[(2.32) \text{those} \quad \text{two} \quad \text{splendid} \quad \text{old} \quad \text{electric} \quad \text{trains} \]

Deictic Numerative Epithet\(_1\) Epithet\(_2\) Classifier Thing

The head of the structure identifies a class of Things (trains), which are denoted by a nominal group, but there may be a lexical extension of the class specification using Classifiers: electric trains. The Deictic element indicates which subset of the Thing is intended (those, the, my, a, each, neither). Epithets can express experiential properties (old) and attitudes (splendid), though there is no hard and fast dividing line between them.\(^7\)

Deictic and Epithet\(_1\) elements have little relation to the ideational meta-function, i.e. they typically do not refer to concepts or in other ways contribute to the representation proper. Other elements combine ideational and interpersonal components in various degrees. As for Numeratives, the author may choose \(\emptyset\) (no specification), few, several, a fair amount, an insignificant amount, a wee bit, twenty, etc, for rendering the amount of objects, which are considered in the same conceptual event. Classifier and Thing can be used to subsume the object being described under a particular subclass, for example, for derogative or admiration purposes, so that the reader is influenced in a way intended by the author.

Epithet\(_1\) elements offer the most immediate way for expressing the authors attitude and influencing the reader. The following example from W.M. Thackeray’s Vanity Fair contains a sequence of evaluations using a single word good:

\[(2.33) \text{yet, as it sometimes happens that a person departs this life who is really deserving of all the praises the stone cutter carves over his bones; who IS a good Christian, a good parent, child, wife, or husband.} \]

\(^7\)Classifiers are also close to properties described by Epithets, but Classifiers and Things together constitute a term for designating a class of objects. Thus, a Classifier does not necessarily designate a property, for example, red wine is not really red in terms of its colour.
2.4. COMMUNICATION-CENTRED DESCRIPTION IN ACTION

The interpersonal contribution of the sentence is related to praising the
person in question for being above a certain standard, even though in the
slightly ironic context, as the author starts with the assumption of the tra-
ditional hypocrisy of epitaphs, admitting that they might be true in some
cases. The ideational contribution of this element is close to zero: it has
very little to say about representational qualities of a person that deserves
the praise. The relationship between good and Christian or parent is not
inherently encoded in the semantics of good, but English allows the evalua-
tion of various qualities using good alone. The possibility to use the generic
good is not always true in English and this does not hold in translations.
The following examples correspond to three possible translations of (2.33)
to Russian:

(2.34) byl xoroshi, xristianinom, xoroshi, roditelem, rebenkom, ...
      was good Christian, good parent, child, ...

(2.35) byl istinnym xristianinom, ljubjašim roditelem, rebenkom,
      was devout Christian, loving parent, child,
      ženoj ili mužem
      wife or husband

(2.36) byl primernym xristianinom, predannym roditelem, ljubjašim
      was exemplary Christian, devoted parent, loving
      chadom, suprugoj ili suprugom
      child, wife or husband

Example (2.34), which uses xoroshij, the direct translation equivalent
of good, is barely acceptable, because of the restricted collocation ability of
Russian nouns in the example with xoroshij.\(^8\) Ethymologically xoroshij is
related to ‘something that looks good’; this is the only sense recorded for
xoroshij in the 11-15 centuries (Varbot, 1980). In modern Russian this word
is normally used for the evaluation of both external and internal qualities,
but only those that are relatively easy to compare against a standard, so
that they can be treated as visible. For instance, it frequently collocates
with such words as cena (price), novosti (news), otnoshenija (relations),
resultat (result), as well as physical objects, e.g. avtomobil’ (car), doroga
(road), kvartira (flat), etc.

\(^8\)Note also that good and bad are treated in NSM as semantic primes that share the
same meaning in all human languages. Yet this example makes it clear that good ≠ xoroshij,
in the same way as anger ≠ gnex.
At the same time the original sense of *good* in Old English was ‘fitting’, ‘suitable’ (OED, 1989), thus, inherently implying the functional evaluation of an object for a particular purpose. Incidentally, the etymological information in (OED, 1989) refers to its Russian cognate, *godnyj*. Diachronic considerations are not always relevant for synchronic descriptions. Many types of evaluations expressed by *good* in English can be naturally translated by *xoroshij* in Russian, e.g. *look, news, quality, reputation*. However, regular patterns of uses of a word that are inherited from an earlier stage of language development put natural restrictions on the collocational abilities of this word in modern language, as it is the case in (2.33).

Even though such expressions as *xoroshij xristianin, roditel’, rebenok* (good Christian, parent, child) are possible, they are not always appropriate translation equivalents to *good Christian, parent, child* in English. In Russian they imply the possibility of a straightforward judgement on the basis of one’s behaviour. For instance, *xoroshij xristianin* attends the church frequently, *xoroshij rebenok* behaves well or studies diligently. Also, *xoroshij xristianin* can be used to distinguish a good Christian from another person, who is *ploxoj xristianin* (bad Christian), making the assumption that the difference between the two is self-evident. Since an epitaph evaluates inner qualities of a person buried under the gravestone and no reference to *ploxoj xristianin* can be used here to make the judgement easily, it is problematic in Russian to use the expression like (2.34).

The second translation (2.35) interprets the relationship between the noun and its evaluation by making the the praise of personal qualities more explicit in ideational terms. At the same time, (2.35) keeps the same structure of the nominal groups as the English original. The quality of a good Christian is praised separately (*a good=devout Christian*), but the qualities of family values are generalised (*a good=loving parent, child, etc*).

Example (2.36) is taken from a real translation of *Vanity Fair* (made by M.Djakonov). It goes further in assigning a specific quality to *parent* — *predannyj* (devoted) and leaving *ljubjaschij* (loving) to appraise the good qualities of *child, wife or husband*. In a Chinese translation of the same text, as reported by Zhang Meifang, the translator interpreted all qualities as a devout Christian, a loving parent, an obedient child, a virtuous wife, or a responsible husband, thus adding the ideational component to every

---

9 There is another difference between the two translations. (2.36) uses nouns that do not fit the current context: *chado* (an archaic word for child) or *supruga/suprug* (a formal word for wife/husband, it is not expected here in modern Russian, but was used in this way in the 19th century). This achieves the rhetorical effect of an antiquated language reflecting Victorian English (the translation itself is relatively modern).
relationship and importing the Confucian values into the translation: a good child is obedient, a good wife is virtuous, and a husband is responsible (it would be unusual to see such praises on a Victorian tombstone).

Even though the Deictic elements in the functional structure of the nominal group normally carry meanings from the textual metafunction, they can also aim at influencing the reader. The following piece, which is written by a death-penalty proponent:

(2.37) Some people ... start to cry about the person being put to death, yet they don’t shed one tear for the person who is really deserving of their tears, the victims.10

contains, among others, two nominal groups referring to the opposition of two persons: the person that is put to death and the victim. The second person is described explicitly with an emotional charge (as a victim and as a person deserving of tears), while the first person is not described as “the criminal” or “the convict”. Rather, the sentence implicitly suggests that the first person is the causer of suffers by means of putting him/her into the adversative relation to the victim. Thus, properties of the first person are underspecified, but s/he is endowed with negative connotations. At the same time, the author of (2.37) does not explicitly claim that the first referred person did the crime, because the argument of the opposite side is that the person charged with the death sentence is, in fact, innocent.

Another reader’s influence in (2.37) is hidden in the Deictic element of some people, which defines a subset without an explicit specification of conditions for belonging to the subset. At the same time, some people is the grammatical subject of the sentence, so the statement in (2.37) does not allow the possibility to affirm or deny the validity of the complete proposition by using the reference to an unspecified group.11 The interpersonal potential behind the modern use of some people may imply the relative unimportance of the group referred in this way. In the context of this statement, this means that the views of the group are marginal and should be dismissed as absurd. The absurdness of their position is further amplified by the author of the fragment using the alleged contradiction in their views: they have compassion for the criminal, not for the victim. As Agre notes, such near-tautological statements as (2.37) are common in the recent political jargon.

11See the functions of the grammatical subject in English according to Halliday (1985:76ff).
in the USA. From the position of cool reasoning, they may have little sense, however, they achieve their function, which consists in persuading others or, more likely, maintaining the intersubjective realm of ideas and arguments for the death penalty and in seeking support within the community of death-penalty proponents.

Naming can be applied not only to natural-kind objects, but to their properties or abstract entities. Properties correspond to physically detectable features (like natural objects they allow an ostensive reference), and, at the same time, they are related to abstract concepts. For example, the adjective *old* in (2.32) refers to the concept of age and/or fashion, while *splendid* to the concept of beauty. What makes an electric train old depends on the rate of changes in railroad engineering and on the experience of the speaker with these changes. Anyway, this expression is not a direct reference to the abstract concept of the age of electric trains, as the expression does not restrict the age of the train. A splendid train is also an inexact reference, which expresses an admiration of the speaker with the trains in view, though the speaker can provide no definition on which trains are treated as splendid.

### 2.4.2 Case Study: terminology in software manuals

Approaches to terminology also follow the two paradigms in lexical research. The logical approach assumes existence of a language-independent ontology of notions each of which can be mapped into a term (preferably one term for one language). Wüster (1979) is a clear example of approaches of this type. On the other hand, the communication-centred approach treats terms with respect to their function in transmitting experience, for instance, the creation of terms as virtual entities (Halliday, 1998).

Terms in a domain exist as a resource for accomplishing transmission of notions, which are in their turn maintained using this transmission activity. This dualism of knowledge transmission calls for treatment of two aspects of terms: formation of their semiotic system and categorisation of a problem domain imposed by the established system. The logic-centred approach means that terms are mere labels for notions in the language-independent ontology of a problem domain. This treatment is satisfactory for a long-established problem domain as, for example, classical mechanics or anatomy, for which notions have been shaped by generations of researchers and practitioners, and now knowledge transfer is re-produced by the system of education. However, even in such domains we can study specific linguistic practices involved in knowledge transfer, see the study of the language of
2.4. COMMUNICATION-CENTRED DESCRIPTION IN ACTION

mathematics in Chapter 11 of (Halliday, 1978). At the same time, in any area subjected to recent rapid developments, such as computer software, new notions and systems of notions, such as principles of a particular graphical interface, are regularly invented and they need terms for their expression in order to be transferred in two types of communities: professionals and naive users.

As shown in (Sharoff and Sokolova, 1996) the register of software employs particularly active use of metaphoric transfer of everyday notions, for example, editor, ‘drag and drop’, folder, mouse, window, in contrast to the use of neologisms by formation of complex words, borrowings or creations from Greek and Latin elements, which are normal in other registers of science and technology. Although this register is mostly monolingual in its origin (the majority of inventions were introduced in the English speaking community first and translated to other languages afterwards), the same trend is valid in other languages, at least, in French, German and Russian, in which official manuals prefer metaphoric translation equivalents for English terms, while the community of computer science professionals (including, but not exclusive to the hacker slang) prefers to borrow original English terms with corresponding morphological alternations.

For some terms, such as folder or drag-and-drop, the metaphorical treatment is natural to their everyday usage, while metaphors implied by editor or mouse are less clearly accessible. In any case of use of a semiotic system, metaphors lie at the heart of constitution of new meanings (Ricoeur, 1978), but regular usage patterns make these metaphors dormant. Later developments in the register call for new metaphors, which partly agree (for example, messages in the folder) or contradict earlier ones, as in a browser is an application, a person looking into a shop window is not like an application of any sort. However, this does transmission of knowledge no harm, since before the term browser appeared, application had already been established in this register as a user-end task applying to system tasks.

In cases of linguistic and cultural differences caused by translation between languages, we find mismatches between metaphors. This happens particularly often in the register of software with its monolingual origin, so software terms in other languages follow metaphors imposed by English. For example, the Russian term priloženie exists as a direct translation of application, however, the Russian semantic system lacks the possibility to treat prilagat’ (to apply) as a request, in this case, from an application to functions of the operating system. The verb prilagat’ implies an activity that extends over the Actee of the verb, e.g. prilagat’ usilie, to apply force, without the possibility of a reference to requesting, which is possible for
apply in English. Another example of a mismatch between the metaphor implied in the source language and its most natural translation equivalent is the Russian verb *podderzhivat’*, which is a translation equivalent of the English *to support*. It corresponds well to such senses of *support* as “bear, strengthen, help, finance”, however, the English idiom *to support software* induced the new Russian usage pattern: *podderzhivat’ programmnoe obozrenie* (in more standard Russian this would mean roughly “to protect software against a physical fall or money shortage”). The term first established as a computer industry jargon and then extended over the wider users community after a short period of resistance from Russian language purists.

These are conditions on formation of terms according to their propositional content (or, in terms of SFL, in terms of the ideational metafunction), however, considerations of the user model (the interpersonal metafunction in SFL) and textual realisation (the textual metafunction in SFL) also influence the lexical choice. Interpersonal influence on terminology involves connotations and stylistic restrictions and expression of a personal attitude for a signified notion; a typical example is the choice of terms depending on the addressee, e.g. a professional, an averaged user in the official style, or a computer-illiterate person.

Translation of English terms also raises the question of additional connotations for their translation equivalents. For example, a Russian manual from 1989 uses

(2.38) *vyberite otvet s pomoshchju manipulyatora “mysh”*

*choose the answer using the “mouse” manipulator*

since the term *mouse* as the input device seemed to be too informal for a written text. This is the reason, why the author uses it in this example as a label which is put in quotes and is classified with a general more scientific term (*manipulator*). At the same time, modern manuals use direct translation: *nazhat’ levuyu knopku myshi* (click the left mouse button. *Icon* is another term with strong interpersonal connotations, which direct translation equivalent in Russian *ikona* is endowed with a religious flavour (icons are venerated as sacred in the Eastern Christian Church), so the term is rendered as either *piktogramma* (pictogram) or *znachok* (small sign).

At the rank of nominal groups, interpersonal influence mostly involves evaluations of terms and their connotations. The interpersonal metafunction plays more important role at the clause rank, in which it provides resources for exchanging meanings in discourse. The most important goal of texts in the register of software is to instruct the reader in ways to perform some tasks. Typically user’s actions are expressed using either:
personal imperative instructions (\textit{Enter ? to display a style list}) or
non-personal statements (\textit{A style list is displayed by pressing ?}).

The choice of non-personal statements is particularly characteristic for original Russian manuals, which tend to describe not instructions for the user, but to present a description of an artificial world of software with its own laws and possibilities for actions. The choice of the personal vs. non-personal style also leads to differences in the rhetorical organisation of a manual. Personal instructions, which are realised by imperative clauses, express commands to be followed by the user, so the text is developed as a sequence of operations to be performed to achieve a task:

\begin{enumerate}
\item \textit{Enter ?} in order to display a style list.
\end{enumerate}

A manual in the non-personal style is developed not through the sequence of actions and corresponding purposes, but through elaboration Mann and Thompson (1988):

\begin{enumerate}
\item Pressing ? displays the style list.
\end{enumerate}

The non-personal style also tends to express processes using nominalisations (\textit{nazhatie, pressing}) instead of verbs (\textit{nazhat’, to press}). This de-emphasises the Actor, which role is also omitted in the nominalised process: the implicit \textit{you} in (2.39) has no (even implicit) position in the structure of (2.40), only the Actee of a nominalised process can be expressed (by the genitive case). Another reason for the wider use of nominalisations in the non-personal style of software manuals is that nominalised expressions are natural in other registers of science and technology, so they are naturally transferred into the domain of software manuals. At the same time the wider use of nominalisations in science owes to the same structural reason as the structure change in example (2.40), i.e. a number of Actees can be expressed using the recursive structure of a nominal group, for instance, \textit{the exchange rate calculation method of adjustment}, while the number of participants in clause-rank configurations is restricted (Halliday, 1978).

However, the choice of the non-personal style is potentially detrimental, because it also distorts one feature that plays the more important role in
50 CHAPTER 2. THE RELATIONSHIP BETWEEN WORDS AND CONCEPTS

the register of software manuals in comparison to other registers of science and technology. The user interacts with software in much the same way as s/he interacts with other people. This allows a metaphoric interpretation of instruments for doing actions (for example, software applications or commands) as real doers, so user actions can be treated as a dialogue with another participant:

(2.41)  *Word may prompt you for a file name*...
(2.42)  *Whenever an application needs Internet access*...

Since notions should be accommodated to the linguistic context in the deployment of texts, words for their expression follow requirements imposed by narrative cohesion, most notably involved in realisation of referring nominal groups, for example, a complete realisation of a concept: the *Multiline Styles dialog box*, a generic realisation: the *dialog box*, a shortened generic realisation: the *box*, a synonymic expression: the *information panel*, an anaphoric realisation: *it*. An example of register-inspired constraints on textual realisation of notions is legally obligatory use of extended nominal sentences in patent claims, as discussed in (Sheremetyeva and Nirenburg, 1996).
Chapter 3

The Form-Meaning interface

3.1 Principles of the corpus-based study

The attitude towards corpus studies is another instantiation of the difference between rationalist and empiricist approaches in cognitive sciences. If an approach is rational, then the evidence is based on linguistic intuition, and there is no need to use (often messy) data from corpora, because the study is mostly aimed at the well-formedness of linguistic structures and the distinction between acceptable and unacceptable sentences can be done by any competent speaker of a language in study. This position is explicitly advocated by Chomsky, who claims that the subject of linguistics is the study of linguistic competence of a native speaker in a homogenous linguistic community, rather than the study of performance (Chomsky, 1965). In his later writings, e.g. (Chomsky, 1987), Chomsky elaborated the distinction between competence and performance by distinguishing between I-language and E-language. Internalised I-language is a mental phenomenon to be studied by linguistics on the basis of logical analysis of linguistic structures in various languages. Externalised E-language, which can be studied on the basis of observable phenomena, is claimed to be dissimilar from I-language and cannot provide the basis for studying the alleged universal structure underlying all human languages. In the case of studies in lexical semantics, the linguist refers to his/her intuition as the ideal speaker that embodies the ideal dictionary, to which WordNet or a set of NSM definitions provide an approximation.

On the contrary, empiricist approaches begin with empirical data. The object of the study is naturally occurring discourse (either spoken or written), which is treated as an instantiation of the system of resources for
CHAPTER 3. THE FORM-MEANING INTERFACE

communicating speaker’s needs. Therefore, such study focuses on aspects of linguistic performance explicitly manifested in linguistic behaviour. As Leech (1992) remarks, a corpus of authentically occurring discourse was before 1950s the thing that linguists meant to be studying. For instance, the activity of collecting and maintaining corpora was an important part of work of lexicographers starting from first monolingual (English and French) dictionaries in the 17th century. The same interest was important for ethnographical and anthropological research, in particular, for Malinowski and Whorf, etc. However, with the advent of the Chomskyan paradigm, the linguistic research based on authentic texts went to sleep for twenty years, like Rip Van Winkle, and gradually woke up, this time under the name of computer corpus linguistics (Leech, 1992:105).\footnote{Of course, the parallelism between the communication-centred paradigm and corpus-based studies is not complete. Many studies in formal linguistics are aimed at studying corpora, see, in particular, development of tree banks, e.g. PennTreeBank (Marcus et al., 1993). Some studies oriented to studying communication acts are not based on authentic discourse, but on invented examples, e.g. analyses by Grice or Searle.}

According to (Sinclair, 1991), corpora provide “hard measurable evidence” which reflects genuine uses of language in ordinary communication. Nowadays huge amounts of texts are available in electronic form. Thus, when a corpus is treated as any collection of texts, corpora are abundant. Even Internet as a collection of texts can be considered to be a corpus (Kilgarriff and Grefenstette, 2003).

However, corpora that are used for linguistic research should conform to a set of minimal requirements, such as discussed by (Sinclair, 1996):

1. quantity: the corpus should be as large as possible;
2. quality: the corpus should be authentic and representative;
3. annotation: a certain level of linguistic annotation of raw texts is desirable;
4. documentation: texts and the annotation scheme should be explicitly described.

What corpus can be considered as large depends on available technologies. For instance, the first large-scale corpus compilation result in computer corpus linguistics is the Brown Corpus (Francis and Kučera, 1967), which sets the corpus of one million words (1 MW) as the target size. However, the corpus now can be considered as rather small, because the standard size for
a modern reference corpus of a language is 50-100 MW, while largest modern corpora reach the size of 1 billion words (Cieri and Liberman, 2002). With respect to the quality, a corpus is a representative collection of authentic texts produced in natural communication without any intervention of the corpus developer. This also calls for the balance of genres in the intended coverage in order to attain some sort of representativeness: a reference corpus must cover all major registers of language use in a proportion that resembles their use by language speakers. Out of this reason Internet cannot substitute a representative corpus. It qualifies to be a corpus in terms of its size and authenticity of texts. For instance, the amount of Internet documents that are indexed and can be found by a search engine can be estimated at about 3 terawords, much larger than any conceivable corpus.

However, there are three types of problems that hinder its use for corpus studies. First, it cannot be claimed that the material is representative and that there is a balance of text types. Texts presented on the Internet are chaotic: their set depends on preferences and interests of a very specific group of people that are active on the Internet. The recall of search results also cannot be evaluated, because it depends on unknown parameters: which texts are available or not available on the Internet; which texts available on the Internet were not found by the search engine used for the query, etc. Second, search engines address the needs of information retrieval, rather than linguistic search. For instance, the title and the keywords field of an HTML document are not good examples of normal word uses, however, they are typically foregrounded by search engines. What is more important for linguistic purposes, search engines cannot restrict morphological properties in queries. Third, search engines present search results in a way that also does not correspond to the needs of a linguist. The pages are ordered in terms of their information rank that has nothing to do with linguistic criteria. The output also does not form a concordance, because pages in the output are separated by documents, rather than by contexts of their uses. Finally, search results are based on words occurring in titles of pages or keywords or even in other pages that refer to the link being displayed as relevant.

The size is estimated as 3 billion pages indexed by Google at the time of writing with the average page size of about 7 KB, i.e. about 1000 words per page in average (Lawrence and Giles, 1999).

See, however, attempts to develop search engines oriented to corpus linguistics, (Kilgarriff and Grefenstette, 2003; Kehoe and Renouf, 2002).
CHAPTER 3. THE FORM-MEANING INTERFACE

3.2 What we can find in corpora

The top-down stance in communication studies starts with what are meanings or communicative intentions and how they are realised in texts as forms. This is the primary stance: corpora are collections of authentic texts produced to realise certain communicative intentions. The process of their production is also top-down: the authors start with meaning intentions and realise them into word forms.

However, when we work with a corpus, the bottom-up stance, which starts from word forms and tries to get their meanings, takes precedence, because a corpus is essentially a collection of forms. One can have access to enumeration of word forms used in the corpus, but not to meanings expressed by the corpus. More specifically, access to the corpus is typically based on the orthographic form of words. There are difficulties with searching for phonetic-relevant information in speech corpora.

The theoretical assumption of the existence of a direct mapping from context to meanings has not only theoretical objects. It encounters practical difficulties even in simplest cases, for instance, in searching for orthographic forms, which use characters other than the 27 basic Latin characters. For instance, in many texts German words with umlauts are expressed using the Windows encoding, which is different from the ISO or Unicode encodings used in other texts, but, eventually, specific conventions for dealing with such characters can also be used, for instance, ı can be encoded as ue, ü or \"u. The situation is yet worse for such languages as Arabic, Japanese or Russian, in which all characters are non-Latin. This leads to development of special encodings, i.e. mappings between internal representation of characters in the computer and their appearance on the screen, so that the computer can find a character on the basis of its encoding and display it appropriately. For instance there are several competing encodings for Cyrillic characters, including DOS, CP866, ISO-8859-5, KOI-8, Mac, Windows (aka CP1251), as well as several representations in Unicode. Each of the encodings uses different values for representing the basic set of Cyrillic characters. For instance, the Cyrillic letter ж (zh) is encoded using the numerical value of 166 in CP866, 230 in CP1251 and 1078 in Unicode. How can we get out of the labyrinth of form-based difficulties in order to get some interesting information from corpora?

Then we have the next layer of problems. The majority of simple applications aimed at working with corpora use queries, which are also based on word forms, so the user can find examples for using a specific word form, like go, or can extend the query by including a list of word forms (go or
3.2. WHAT WE CAN FIND IN CORPORA

Even enumeration of words used in the corpus is a difficult enterprise, because the detection of words out of word forms is often ambiguous, e.g. considering whether *left* is an adjective or a past verb form. The solution of complete form enumeration is cumbersome, but even if it is feasible for English, it is definitely too complicated for highly inflecting languages, for example, in Russian a verb may have more than 30 different forms, many of which cannot be found using simple wildcard queries, e.g. the following are forms of just one verb *идти, иду, идет, шел, шёл, шёл, шёл, шлет* etc.

Since the search procedure is based on forms and research interests are based on intended meanings, the most problematic cases concern situations, when the user of a corpus searches for contextually dependent properties. For instance, it is hard to detect situations, in which a motion verb refers to non-motion events, like *The inauguration went well*. Similarly, if the user is interested in all expressions of emotions in a text, the search should not be restricted to labels of emotional states, like *disgust*; it should also include indirect hints (like *frown*) or evaluations (like *awful*), which are reliable indications that an emotional state is expressed. Similar problems concern the use of corpora for studying grammatical phenomena. As (Halliday, 1992) remarks, no program can retrieve all clauses with marked circumstantial theme.

The problem has also a straightforward solution: the corpus developer should add as many annotations to word uses as necessary. Possible tags can include lemmas, parts of speech, syntactic functions, semantic interpretations, etc. This is the technique adopted in the corpora discussed in previous chapters. However, the procedure of adding annotations automatically is also based on word forms and not on meanings intended by their uses, so it is prone to errors in meaning detection. At the same time, other tags can be added using semi-automatic procedures, like alignment between sentences and words of parallel texts. Some more tags can be added manually, like the distinction between physical and imaginary motion. However, manual or even semi-automatic operations make corpus preparation much harder and more expensive, so heavily annotated corpora happen to be smaller in size and, hence, less representative (this is also the reason for the smaller size of the aligned corpora used in the study).

Even the definition of what is a word and what is a sentence may present a problem: the notion of what is a word form is already a meaning assigned to what is a token in a sequence of characters in the text. A simple definition of a word form as a sequence of characters between spaces and punctuation marks does not work. Do we ask the computer consider a hyphen as a word
CHAPTER 3. THE FORM-MEANING INTERFACE

separator, as in *co-ordinator*, *morphine-like* or in *43-year-old?* How about the apostrophe? How many words and what are those words in *can’t, I’ll or last week’s show?* For a comprehensive discussion of technical problems and potential solutions for tokenisation of free-running texts, see (Grefenstette and Tapanainen, 1994). On the other hand, a sequence of several word forms may constitute a multiword. For instance, there is no sensible function to assign to every separate token in *as well as in English or po mere togo kak* (in the process of) in Russian, so it is useful to consider them as single words, but as Kilgarriff (1997b) notices, this limits the options for studying words which other uses are related to multiword ones. For instance, *no doubt* can be considered as a multiword unit, but then the searches for *doubt* will have no context for *no*.

An automatic extension of word forms for a lemma is also feasible, however, it is much less reliable then the detection of word forms: it is not always easy get from a word form to lemma partly because of the inherent ambiguity that exist in any natural language (see the analysis of the Russian word frequency list in Section 3.4 below), and partly because of theoretical problems concerning the identification of a lemma. Typical problems concern the decisions on treating participles as forms of verbs or adjectives (*perceived* or *unwanted*, there is no verb in the latter case), reflexive verb forms as separate from non-reflexive ones in Russian (*osušestvljat’,* to perform vs. *osušestvljat’*sja, to be realised), the detection of verbs with separable prefixes in German (*zurückbringen*, to bring back), etc.

Some researchers, for instance, John Sinclair (1991), treat the identification of word and lemma as not useful for corpus-driven research, because quite frequently different word forms perform different functions. Sinclair suggests starting an investigation with individual word forms, which may be combined into a single lemma, if the variety of forms is supported by corpus evidence. For instance, the noun *yield* in the sense “agricultural produce” as in *low average yields of half a ton of rice per acre* is prominently associated with the plural form (Sinclair, 1988:76). At the same time, for flective languages, such as Russian, the number of forms available for a word prevents us from studying word uses on the basis of specific forms. For instance, every adjective in Russian inflects for the case, gender and number, giving about 15 forms for every adjective. Even if the gender can be a potentially useful distinction, no semantic difference can attributed to the variation in the case and number:

(3.1) na prudu byl belyj lebed’ vs.
on pool was white-nom swan
3.2. WHAT WE CAN FIND IN CORPORA

There was a white swan in the pool.

(3.2) na prudu ne bylo belogo lebedja
    on pool was no white-gen swan

There was no white swan in the pool.

Another problem that concerns corpus studies is that corpora can only provide the positive service: corpora can conform but they cannot deny. Concordance lines can provide the evidence for intuitive understanding of a certain phenomenon. They can also help to find linguistic constructions which cannot be taken from intuition or to correct the intuition with respect to allegedly impossible examples, which can be actually found in corpora. At the same time, the negative service of corpora is possible only indirectly: the absence of certain constructions in the corpus does not reject the possibility of their actual use in language, because our linguistic intuition can provide quite plausible examples which are not found in the limited corpus, but still exist in everyday language. If aftertaste\(^4\) occurs 19 times in the BNC and only one context is positive (others are bitter, nasty, acidic, half-digested, etc), does this tell EFL students that pleasant aftertaste is bad English? Even though there is no single instance of pleasant aftertaste in the BNC, the phrase is perfectly normal and in a larger corpus, for instance, in the Internet we can find dozens of examples of exactly this phrase (mostly in the description of coffee and wines). Also out of 19 uses of aftertaste 12 occur in the clause-final position:

(3.3) something went wrong producing an unsatisfying dish with a bitter aftertaste.

Is this fact important for its use? On the one hand, yes. The clause-final position is associated with a regular pattern of uses: X has an aftertaste or X with an aftertaste, i.e. this is a property attributed to an object that causes a certain feeling after an event (digestion or otherwise, as in 3.5 below), and attributes in relational clauses in English (Halliday, 1985:Ch. 5.4) are typically expressed at the end of the clause (even though the realisation using X with an aftertaste is not a congruent clause). But on the other hand, the answer is no. The cases, when the word aftertaste does not occur in the clause-final position are legible even within the “relational-clause” framework and can be explained by other factors, such as the discourse development pattern (3.4) or the rhetorical structure (3.5):

\(^4\)The example comes from a private communication with Dominic Stewart.
(3.4) Despite the somewhat vinegary aftertaste we found it very easy to drink.

(3.5) on those rare occasions when it [success] comes we are always left with a nasty aftertaste to spoil it.

As we see, the problems with using corpora are formidable. At the same time, no source is better for studying regular patterns of word uses than a large representative corpus. Fortunately, computer corpora do not only cause problems in finding information, they also offer solutions by giving evidence that is hard or impossible to obtain from other sources. This concerns detection of collocations and word similarity classes.

A list of concordance lines sorted according to the immediate left or right of search expression helps in detection of regular patterns of uses. For instance, the concordance for the word post shows that its frequent function in English is to introduce scientific concepts, such as graduate, natal, mortem, traumatic, etc. However, it is hard to assess the variety of words associated with post in this way. It is also known that humans are very efficient in judging the possibility of an expression, e.g. whether two words can co-occur together, whereas for the computer the task involves searching through the complete corpus.\(^{5}\) However, humans are much less efficient in listing the most frequent patterns of uses, whereas there are several efficient methods for assessing the strength of relationship between lexical items and presenting the strongest collocates. The present study uses MI-score, t-score and loglikelihood score.

Two words can be considered completely randomly distributed in texts. In statistics, mutual information (MI) is information that one random variable contains about another. Thus, the MI-score reports the degree, in which the joint frequency of two words is greater than the probability of their appearance together, if the two words were indeed distributed randomly (Church, 1991), (Church, 2000). It is formally defined as:

\[
\text{(3.6) } \text{MI}_{ij} = \log_2 \left( \frac{O_{ij}}{E_{ij}} \right); \quad O_{ij} = \frac{F_{ij}}{N}; \quad E_{ij} = \frac{F_i}{N} \times \frac{F_j}{N}
\]

where \(F_i\) is the number of occurrences of each word in the corpus, \(F_{ij}\) is the number of their joint occurrences, \(N\) is the number of words in the corpus, \(O_{ij}\) the frequency of joint occurrences relative to the corpus size, \(E_{ij}\) is the expected frequency of random distribution. The value of the

\(^{5}\)And even after that the results need human interpretation and adjustment, see the example with positive aftertaste.
3.2. WHAT WE CAN FIND IN CORPORAS

MI-score is zero, if the occurrence of one word is completely independent from the occurrence of the other. The MI value is greater than zero, if the two words occur together more frequently than expected. By the nature of the equation, it is biased towards words rarely used separately in the corpus, including occasional uses of rare words. For instance, the most significant MI-score collocates on the left of agreement in the BNC are solus, Schengen, debitor-creditor-supplier, Anglo-Irish, as well as rarities specific to a corpus, like UKUSA, Ho-Sainteny. The most significant collocations with post are indeed scientific terms post facto, post quem, post-modernism, post-infarction (as in patient, angina, etc), as well as post office, box, code, post Cold War (age).

The t-score of a collocate weights the difference between the expectation and co-occurrence according to the joint occurrence frequency, i.e. it pays more attention to the most frequent co-occurrences (Sinclair, 1991). The t-score is defined in the following way:

\[(3.7) \quad T_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{O_{ij}}};\]

The most significant t-score collocates on the left of agreement are function words an, the, this, any as well as reach, general, trade, international, peace, sale, mutual, sign. Their most significant t-score collocates on the right of post also include function words of, in, as, at, as well as office, staff, box, graduate, house.

Collocation lists can also help in detection of the semantic “similarity” between two words (Manning and Schütze, 1999:Ch. 8.5), i.e. the similarity of their lexical behaviour. Two words are said to be similar, if the sets of their collocates are similar. For instance, the words compromise and constitution cannot be listed in a dictionary as synonyms to agreement, however, the word compromise has such collocates as: reach, find, seek, accept, negotiate, while the word constitution collocates with under, amend, violate, revise, according to, which frequently occur in the same contexts as the word agreement. In the end, the list of words contextually similar to agreement are:

arrangement, cease, compromise, constitution, contract, deal, guideline, initiative, legislation, negotiation, ordinance, pact, peace, plan, policy, proposal, regulation, resolution, settlement, treaty.

The list of collocates gives a condensed picture of the variety of uses of a word, which then can be studied in details. For instance, the t-score

---

6The list is based on a database collected by Dekang Lin (1998).
CHAPTER 3. THE FORM-MEANING INTERFACE

colloca
tes of *agreement* identified above suggest the study of contexts in the following groups: a discussion its result, a legal document and a plan for action. Then we can compare them against the t-score collocates of the words *dogovor* and *soglasie* in Russian: *zakljuchat’* (to reach), *podpisyvat’* (to sign), *mirnyj* (peace), *trudovoj* (labour [contract]) for *dogovor*, vs. *davat’* (to express), *polnyj* (complete), *prinicipial’nyj* (in principle), *vza-imyj* (mutual) for *soglasie* which clearly suggests their respective domains in which translations operate. The comparison of collocates for *post* in English and Russian shows that the two words are “false friends” in the majority of contexts, the fact that is obvious for any experienced translator, but not for a novice, as *post* in Russian collocates with *gensek* (the General Secretary), *predsedatel’* (chairman), *gubernator* (governor), as well as *post* GAI/DPS (traffic or patrol police outpost), *post oxrany* (security office), *post upravlenija* (control board).

On the other hand, the class of words whose lexical behaviour is similar to *agreement* mixes its basic types of uses. However, since types of uses can be identified from the collocation list (or from intuition), the similarity class helps in extending the number of patterns, which involve words like *agreement*, because it suggests to test other words and compare their uses in the pattern, when a corpus is too small to give sufficient evidence. Typically, the frequency of trigrams and tetragrams consisting of exact words is very low, even for the corpora of the size of the BNC. For instance, there only nine instances in the BNC of the search expression *close to [0,2 agreement* (the distance between *close to* and *agreement* is not larger than 2 words). This is too small for studying regular patterns of events being close to an agreement on the basis of the BNC, however, if we replace the word *agreement* with its semantic class, we will get 26 concordance lies, which offer more material for the observation.

The studies reported in Part II of the book make heavy use of these tools. Collocates were mostly used for distinguishing types of uses of size adjectives and verbs of motion (these words are polysemic in their nature), while semantic similarity helped in working with emotion words (they are mostly monosemous, which simplifies grouping them according to their contexts of uses).

The following sections briefly introduce the corpora used in the study, including the Russian Reference Corpus, and report the results of two experiments aimed at making sense out of word forms in the corpora: building a word frequency list of modern Russian and studying the “conduit” metaphor and its translations in a corpus of aligned texts.
3.3 Corpora used in the study

The study presented here is based on reference corpora for English and Russian: the British National Corpus (BNC), and the Russian Reference Corpus (RRC). The BNC is a balanced representative corpus resulted from years of careful compilation (Aston and Burnard, 1998); it comprises about 100 million words (MW). RRC comprising about 50 MW has been developed in the course of the study and is partly described below, for a longer description see (Sharoff, 2003b, 2004b). As there is no reference corpus for German, several corpora have been used: COSMAS, IMS and XLEX. COSMAS is the largest German corpus with about 1,600 million words. However, the three German corpora are biased with respect to the coverage of genres: they are mostly based on newspapers. Hence the frequency and patterns of uses of lexical items in them cannot be characteristic for the standard German. For instance, the frequency of occurrences for SPD, a major German party, is comparable to that of gestern (yesterday) in COSMAS, wissen (to know) in IMS, sehen (to see) in XLEX. COSMAS has also a subcorpus, called Mannheimer Korpus (MK). It mostly consists of literary texts, so it cannot be used as the reference comparable to the BNC and RRC; it is also relatively small (its size is less than 3 MW).

Since the size of all corpora differs, reporting the actual number of occurrences of a search term can be misleading: it is not easy to say whether anger with its 3291 instances in the BNC is more frequent than gnev with its 1670 instances in RRC or Arger, which has 20,219 instances in COSMAS and 169 instance in MK. So the reported frequency of lexical items is normalised to the number of their instances per million words (ipm), that is dividing the number of occurrences by the corpus size and multiplying the result by 1,000,000. The measure is useful, because lexical items that are of interest for corpus research have frequencies in the range of 500-10 ipm; for instance, the frequency of anger in the BNC is about 33 ipm, gnev in the RRC—34 ipm, and Arger—13 ipm in COSMAS and 56 in MK. Another useful measure is the frequency band measured in terms of thousands of most frequent words (Kilgarriff, 1997b). For instances, the study of size adjectives reported in Section 5 considers the first 2000 most frequent words, while the study of emotion words (Section 7) considers the first 5000 most

---

7 To facilitate this and similar corpus studies, I have developed some software tools for querying word uses and detecting their most significant collocates, they are available (Sharoff, 2003a). The search interface is uniform for the BNC and RRC, as well as for the parallel corpus considered below. It is modelled on the basis of the IMS Corpus Work Bench (Christ, 1994).
frequent words in respective languages.

With respect to corpus annotation and documentation (other requirements mentioned by Sinclair), the vast majority of corpora developed recently are based on XML (eXtensible Markup Language). The primary reasons for the active use of XML in corpus research are related to the last two basic requirements according to (Sinclair, 1996). In particular, XML allows to separate encoding of the structure, content and presentation of resources. The structure, i.e. the annotation scheme of a corpus, is explicitly encoded by means of document type definitions (DTDs), the content conforms to the DTD, and the presentation is controlled by means of XSLT: a language for transforming XML documents. Availability of DTDs for specific types of documents allows standardization of the structure of documents and the possibility of their reuse in other projects.

There are several competing standards for encoding corpora, including the following: EAGLES (European Advisory Group on Language Engineering Standards), TEI (Text Encoding for Interchange), and XCES (the XML Corpus Encoding Standard). The TEI guidelines (Sperber-McQueen and Burnard, 2002) define basic conventions for tagging texts of different types, including linguistic annotations, though they are not specifically targeted to linguistic tasks. This is also the most widely used standard. (EAGLES, 1996) provide guidelines for dealing morphosyntactic annotation of corpora on a very general level. The four requirements for the corpus design from (Sinclair, 1996) are actually part of the EAGLES guidelines. However, EAGLES do not provide actual DTDs for corpora. XCES (Ide and Romary, 2001) is the most general framework for linguistic annotation of corpora and is aimed to be a language resources management standard for the International Standards Organization (ISO TC37/SC4). It defines an abstract metamodel, which captures all potential models of linguistic annotations, using a hierarchy of data categories for abstract tags <struct> for a node and <feat> for its features. This provides possibilities to encode parallel annotations on different levels, alternative annotations in the case of ambiguities or disagreement between annotators, aggregation of annotations for several (potentially disjoint) linguistic units into one annotation unit, e.g. for verbs with separable prefixes in German, or splitting annotations of a single word, e.g. for zum into zu dem in German. Another aim of XCES is to implement the EAGLES guidelines. However, the generality of the XCES mechanism makes it cumbersome for using in simple applications. The encoding used in the study is based on the TEI framework in terms of DTDs and follows the basic EAGLES guidelines in terms of linguistic annotations.
3.3. CORPORA USED IN THE STUDY

3.3.1 The Russian Reference Corpus

<table>
<thead>
<tr>
<th>Genres covered</th>
<th>Nr of words</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>22.54</td>
<td>45.73</td>
</tr>
<tr>
<td>News</td>
<td>16.42</td>
<td>33.31</td>
</tr>
<tr>
<td>Non-fiction</td>
<td>9.21</td>
<td>18.69</td>
</tr>
<tr>
<td>Science</td>
<td>1.12</td>
<td>2.27</td>
</tr>
</tbody>
</table>

Table 3.1: The composition of the Russian Reference Corpus

At the time of the project, no reference corpus of Russian existed. The only available balanced corpus of Russian was the Uppsala corpus, which had been modeled on the basis of the Brown Corpus for English (1 MW). Thus, it is too small for modern standards. Also, the Uppsala corpus is not freely distributed, has no lemmatisation and possibilities for studying collocations.

This led to development of the Russian Reference Corpus, which contains a selection of modern fiction, non-fiction texts, newspapers, and science (its current size is about 50 million words, MW). The size of the corpus depends on 1) the availability of electronic texts and 2) the balance of genres. Huge amount of Russian texts were available from the Internet. A selection among them provided the balance between the genres and text types roughly according to the model established by the BNC. Currently the Russian Reference Corpus is in the process of further development to achieve better representation of text types Sharoff (2004b). The distribution of texts in the corpus used in the current study is given in Table 3.1.

All texts were written originally in Russian in the period between 1970 and 2002 (newspaper texts from 1998-2001). Translations from other languages were deliberately excluded, because translations often include word choices and grammatical constructions that are not characteristic for original texts. The Russian Reference Corpus also helped in development of the frequency list of Russian discussed below. The corpus has been automatically annotated with part-of-speech (POS) and lemmatised.

3.3.2 The aligned corpus

The contrastive study of language-specific ways for communicating experience should be also based on the empirical evidence provided by parallel corpora, which consist of source texts and their translations to other lan-
CHAPTER 3. THE FORM-MEANING INTERFACE

The availability of several translations of the same text also allows for empirical analysis of paraphrastic possibilities in one language. Unfortunately, in comparison to multiple monolingual corpora, relatively few aligned corpora are available. They are also much smaller and less representative. Aligned corpora are available for the English-French (Canadian Hansards) or English-German pairs, but not for Russian. The task of the contrastive study of lexical semantics in English and Russian led to development of an aligned parallel English-Russian corpus.

The corpus has slightly less than 1 MW and consists of two technical texts:

- Microsoft Word ’97 User Manual and its translation into Russian;
- excerpts from the AutoCAD v.13 User’s Manual (Chapter 2: Drawing objects) and its translation into Russian;

and three literary texts:

- Lewis Carroll’s “Alice’s Adventures in Wonderland” and three of its translations into Russian by Demurova, Nabokov and Zakhoder and one translation into German by Zimmermann;
- Vladimir Nabokov’s ‘The Vane Sisters’ and two of its Russian translations made by Ilyin and Barabtarlo;
- Vladimir Nabokov’s ‘Lolita’ and its Russian translation made by Nabokov himself.

All texts were initially written in English. Only one German text is available in the corpus, so the corpus is mostly restricted to the English-Russian pair.

Alignment of the text pairs on the sentence level has been performed by means of the Marc Alister software package (Paskaleva and Mihov, 1997), which implements the Gale-Church algorithm for language-independent alignment (Gale and Church, 1991). The algorithm restricts itself to the paragraph level alignment, assuming that a paragraph of the translation always corresponds to a paragraph in the source text. Then, it tries to align sentences in each paragraph according to the principle that longer sentences of the source text (the length can be measured in terms of words or characters) tend to be translated by longer sentences in the target texts. Mismatches between the length of sentences are typically caused by the fact that a sentence

---

8The texts were prepared within AGILE, the EU project aimed at multilingual generation of CAD/CAM manuals, cf. (Bateman, et al, 2000).
3.4. THE FREQUENCY LIST FOR MODERN RUSSIAN

The availability of a reference corpus allowed to compile the frequency list of word uses in modern Russian. The result is important for Russian corpus studies in general and vital for the studies reported in Part II, because the comparison of the Russian frequency list to that of English (Kilgarriff, 1996) helps in making claims about similarities and differences in English and Russian expressions in the same domain.

Up to now “Chastotnyj slovarj russkogo jazyka” (Zasorina, 1977) provided the most widely used frequency list for Russian. However, the corpus from the source is split or two source sentences are joined in the translation. The best alignment in such cases is established by means of optimization of sentence length matches using dynamic programming techniques.

The Gale-Church alignment algorithm is simple and efficient, however, it frequently fails to detect translation equivalence between sentences. So, the results of alignment should be corrected manually. The alignment of “Alice in the Wonderland” and respective texts with its German and Russian translations has been corrected to the possible extent; the alignment of other texts has been mostly left unchanged. However, the quality of automatic alignment of other texts was much better, because the Gale-Church algorithm is noise-sensitive. The “noise” in Alice is related to the reported speech, which is coded in the original English and Russian texts quite differently. By this reason, sentence boundaries and their alignment cannot be determined automatically in this case. Anyway, alignment errors that remain in the corpora do not cause significant problems when searching for translation equivalents, because the consultation function (see below) may present a wide context for each occurrence of a lexical item.

In spite of the fact that “Alice in Wonderland” is a relatively short text (34,359 words) that does not allow to make statistically valid claims about word uses and that is written in a figurative language, the advantage of using it for giving examples is the possibility to establish a larger context for every example, because the text itself is readily available. Even its small size helps in observing all relevant occurrences and studying phenomena that cannot be studied effectively in a larger corpus, such as the “conduit metaphor”, see Section 3.5 below. Finally, the corpus used in the study has three translations of ‘Alice’ to Russian, which helps in the empirical study of paraphrases, when the same situation is reported differently in various translations.

3.4 The frequency list for modern Russian
used in Zasorina is relatively small according to modern standards (about 1 million words). It is outdated: mostly it covers uses from 1920s to 1960s and includes a high proportion of ideological sources, like texts by Lenin and Krushchev and Soviet newspapers, thus, word frequencies in it are severely biased. For instance, sovetski (Soviet) and tovarisch (comrade) are in the first hundred of Russian words on a par with function words. Another relatively recent frequency list for Russian (Lönnengren, 1993) is produced on the basis of the Uppsala corpus. Thus, it is also based on a small population of texts (see the section on the Russian Reference Corpus) and has much more restricted distribution in comparison to (Zasorina, 1977). Finally, both lists are not available electronically.

In addition to the practical importance of a Russian frequency list for the studies reported below, the very process of its compilation reveals some important facts about the form-meaning interface in corpus linguistics. Making a frequency list for any language involves problems of two types. First, many word forms can have different morphological analyses, including several lemmas possible for a word form. Such ambiguities should be somehow resolved in the process of making a frequency dictionary. Section 3.4.1 reports problems specific to Russian.

Second, word forms are not uniformly distributed in texts, so some of them may come to the top of the frequency list completely accidentally, simply because texts of a particular type are available in the corpus. Kilgarriff (1997b) refers to this as the “whelks” problem: if a corpus contains a text about whelks, the word whelk may be unproportionally frequent in the frequency list. Section 3.4.2 describes methods for dealing with the problem in Russian.

3.4.1 Resolving the ambiguity of forms

The diversity of Russian morphology gives the impression that cases of ambiguity, at least, the ambiguity of lemmas corresponding to forms, are relatively rare. Parts of speech have distinct morphological structure, e.g. a verb is structurally different from a noun, so the cases similar to English can, filter seem to be relatively rare.

However, an attempt to build a frequency dictionary shows that about 25% of word forms in texts can potentially be instances of several lemmas, partly because some rare cases concern very frequent words, for instance, byli (be-past.pl vs. story-gen.sg), dorogoj (road-inst.sg vs. dear-nom-sg-masc), stali (become-past-pl vs. steel-gen-sg), tri (three vs. rub-imper-sg), uzhe (already vs. narrow-compar vs. grass-snake-loc-sg). There are also cases
of the regular ambiguity, for instance, personal pronouns vs. possessive pronouns (ego, his vs. he vs. it), short adjectives vs. adverbs (prekrasno, beautiful), adjectives vs. deadjectival nouns (znakomaja, known vs. friend-fem). In most cases, the complete syntactic parsing is required for precise detection of lemmas intended by the author. However, many cases of the ambiguity can be resolved on the basis of statistical criteria of two types:

1. frequency of a lemma. For instance, byli, meril, stanu, zabral are forms of frequent verbs and rare nouns, so a verb (byt’, merit’, stat’, zabrat’) is selected as lemma in such cases.

2. frequency of a word form relative to other forms. For such forms as pora, stali both lemmas are frequent. However, since they constitute the majority of uses of the noun stalj and all uses of the adverbial pora, respective lemmas are selected (and not the verb statj and the noun pora, respectively). Otherwise, the target lemma would not be represented in the frequency list at all.

However, there are cases of genuine ambiguity, in which no decision can be made using an automatic procedure, for instance, vinu, vina, both of which can correspond to two nouns: vina (guilt) and vino (wine). Such analyses are stored as alternatives in the corpus and the selection of forms for the frequency list in the case of genuine ambiguities is random.

The following is a classification of some important cases of ambiguities, in which the decision for selection of a lemma can be made on the basis of the statistical criteria. !Do we need the following for any theoretical or practical interest?

First, many Russian surnames are derived from nouns and can be considered as word forms of respective nouns, for instance, Kuznecov, which can also be analysed as the genitive plural form of kuznec (blacksmith). The same applies to other frequent surnames, such as Orlov – orel (eagle), Popov – pop (priest), Zhukov – zhuk (beetle). The situation is different from English, because the two words in Russian have different lemmas: Kuznecov vs. kuznec (unlike Smith vs. smith), so that we need a proper lemmatisation procedure. Fortunately, the capitalization of the first letter provide the reliable clue for disambiguation. Given names are not typically derived from respective nouns, but they can be eventually homonymous, so the case includes also several given names, e.g. Jakov – jak (yak), Jurka – jurkij (nimble), Tolja – tol’ (roofing felt).

Second, many noun forms allow several analyses, for instance, doka, lozhka, pola, popa, spina. Many nouns also have plural forms which use are
radically different from those of singular forms, so plural forms are considered as separate lemmas, e.g. *den’yi*. The genitive form of some masculine nouns is ambiguous with the nominative form of feminine nouns, for instance, *Evgenija, Zhukova, komponenta*. In some cases, loan words or parts of place names are ambiguous with other lemmas, for instance, *Mak Kinli* (the Russian spelling of McKinley) coincides with *mak* (poppy), *San Francisko* coincides with *san* (rank).

Third, there are ambiguities in detection of lemma for adjectives and nouns derived from them, for instance, *belyj, ljubimaja, masterskaja, polovoj*, as well as in detection of one possible lemma for several adjectives, for instance, *bol’shój/ból’shij, jazykovój/jazykóvýj, vrémennýj/vremennýj*.

Fourth, many adverbs coincide with short forms of adjectives, e.g. *dostupno, neobkhodimo, strashno*, etc. The ambiguity between adverbs and nouns is also possible, for instance, *potom, tselikom, tut*.

Fifth, there are many cases of ambiguity between forms of verbs and nouns, for instance, *beregu, beret, nachal, vesti, vypadu*, as well as between several possible verbal lemmas, e.g. *nashli, poshli, vyshli*.

Finally, there are (relatively few) cases allowing several analyses simultaneously. For instance, *pari* can be analysed as one of two nouns (*par’ı* and *parja*), or the imperative verb form (*paritj*); *uzhe* as the locative case form of the noun *uzh*, as the adverbial *uzhe*, and as the comparative adjectival form of *uzkij; dusha* as one of two nouns (*dush* and *dusha*), or the adverbial participle verb (*dushitj*).

Even though the structure of the Russian frequency list corresponds to the structure of the BNC list and allows easy comparison between the two languages, the units for measuring the length of texts in words, and, hence, the relative frequency of separate words are not completely identical in English and Russian. The following nominal group: *A proper understanding of the history of the surface of the Earth* consisting of 12 words in English is naturally expressed by 5 words in Russian *pravil’noe ponimanie istorii zemnoj poverxnosti*, because of the trivial difference between the structure of the nominal group in the two languages: Russian lacks articles and expresses the relationship between nouns using the genitive case (realised by the ending of the noun) instead of a separate preposition (*of*). Also the shorter length of English texts expressing essentially the same content as in Russian can be caused by the frequent omission of copulas and auxiliaries in Russian and the recursive structure of the verbal group in English (*will have been going*), which is typically rendered lexically in Russian, for instance, the appropriate prefix or aspect. This suggests that the normalised frequency of content words in Russian is slightly higher, and this should be
3.4. THE FREQUENCY LIST FOR MODERN RUSSIAN

... taken into account, when wordlists are compared in Part II of the book.

3.4.2 Resolving problems with word uses

Words are not uniformly distributed in texts. Some of them (like prepositions) occur in many texts with predictable rate, some (like pronouns or mental verbs) are significantly more frequent for certain writers or genres, while some are “contagious”: if a word (e.g. a proper name, a title of nobility or a technical term) occurs once in a text, it tends to be repeated, thus boosting its frequency in a document.

It is widely known that large texts present a problem for frequency lists, since a large text that contains many instances of a rare word can boost its frequency. Large texts are quite frequent in fiction or in news corpora. As an example, the Russian Reference corpus contains a huge sequel to Tolkien’s “The Lord of the Rings” written by a Russian author (Nick Perumov). In spite of the fact that the length of the sequel is about 250 thousand words (kW), less than one percent of the whole corpus, the frequency of uses of the word hobbit in that book puts the word in the first thousand of most frequent Russian words, if no precautions against large texts are made. Out of this reason, the frequency list is calculated under the condition that no single text from the corpus contributes more than 20 kW and no author contributes more than 500 kW to the count. Thus, the subset of the whole corpus used for frequency count is about 27 MW.

However, restrictions on the size of each title and the contribution from each author do not save from boosting frequency of certain lexical items, which are very frequent in books by different authors addressing the same topic. This concern lexical items which frequency is determined by eventual factors, external to language. For instance, the word admiral is relatively infrequent in Russian. However, if a naval officer in the admiral rank is a character in a story, then almost any reference to the character uses the title. If the character is important for the story, there are many references, which inflate the frequency of the word. The “admiral” problem is also characteristic for hereditary titles, names of certain professions (e.g. doctors), technical terms, proper nouns, etc. This feature helps in identification of words important for information retrieval, see (Church and Gale, 1995), but they can significantly distort ranks of lexical items in the frequency list.

The variation in word frequency across texts can be measured in a variety of ways, see (Church and Gale, 1995), (Church, 2000), (Kilgarriff, 1998). For compilation of the frequency list, the easiest way is to use the coefficient of variation, which is defined as the standard deviation divided by
Figure 3.1: The diagram of the frequency distribution
3.4. THE FREQUENCY LIST FOR MODERN RUSSIAN

<table>
<thead>
<tr>
<th>V (in)</th>
<th>Ja (I)</th>
<th>Odin (one)</th>
<th>Admiral</th>
</tr>
</thead>
<tbody>
<tr>
<td>27109.97</td>
<td>17290.39</td>
<td>2889.27</td>
<td>19.41</td>
</tr>
<tr>
<td>7127.69</td>
<td>13020.43</td>
<td>1194.99</td>
<td>182.73</td>
</tr>
<tr>
<td>0.26</td>
<td>0.75</td>
<td>0.41</td>
<td>11.16</td>
</tr>
</tbody>
</table>

Table 3.2: The frequency distribution for some Russian words

The difference in the frequency distribution of different words is quite significant. Three words are studied in Figure 3.1: a preposition, v (in), a pronoun, ja (I) and a content word, admiral (admiral); a cardinal have been also added to the table: Odin (one). Even though the preposition v occurs in different texts with variable frequency, it is present in all texts and the difference between its frequency in different texts is relative low (the coefficient of variation is 0.26). The cardinal number Odin (one) also occurs relatively frequently in many texts (its frequency is 2873 ipm, but the coefficient of variation is 0.41). However, in some texts its frequency is boosted by the fact that is used in specific patterns, like Odin iz X (one of X). This boost is partly related to the author’s style or topic of the message, because it might be caused by not only strictly numerical constructions, but also by superlatives, e.g. Odin iz luchših (one of the best).
In contrast to them, even though the mean frequency of the pronoun *ja* is comparable to that of the preposition *v*, its distribution is much less uniform (this is also indicated by its coefficient of variation 0.75). The figure at the bottom of Figure 3.1 shows that *admiral* occurs in very few texts, but when it occurs, it is very frequent; thus confirming the intuitive expectation of its irregular behaviour. This is also reflected in its coefficient of variation: 11.16. See also its frequency data in specific texts in Table 3.3. However, if we calculate its mean frequency for all texts, including those where it does not occur, the value is relatively high: 19.41 instances per million words (ipm), which brings it to the first 5000 of most frequent lexical items, leaving such words as *ususchestvovatj* (perform), which mean frequency is 19.37 ipm, outside of it. On the other hand, it is too cumbersome to filter such accidental words out of the top list.

For these reasons, the procedure for detecting most frequent Russian words that belong to the core of the language core used an essentially statistical criterion: the coefficient of variation. The value of the coefficient of variation was calculated for the ten thousand most frequent Russian words according to the frequency count over all texts. Words which coefficient of variation is low are generally more frequent in comparison to words, which coefficient of variation, and vice versa. However, many “contagious” words, which do not belong to the core lexicon but are frequent according to the frequency count, are effectively filtered out.

### 3.5 The conduit metaphors in translations

Another example that illustrates the possibilities and problems involved in the study of meanings using “bottom-up” approach is provided by the “conduit metaphor”. Reddy (1979) coined the notion for referring to a regular type of meaning transfer in the

English grammar, which consists in reinterpretation of verbal expressions in terms of physical process of object transmission. Actually the metaphor is related to the investigation of cryptotypes, in particular, the notion of “containers” in the Whorfian Standard Average European (SAE) model.

According to the conduit metaphor, meanings are treated as packaged in words, sentences and texts, as in containers which hold a content and are transferred from the speaker to the addressee. Modifications of the conduit metaphor concern expressions, which describe how the content is arranged in the container and how the addresssee accesses it. As it is impossible to search for metaphors in a text without special annotations, we have to
3.5. THE CONDUIT METAPHORS IN TRANSLATIONS

search for some indications of the conduit metaphor. The search used the following algorithm: search for collocations of verbs that can the transfer or its result, e.g. carry, get, find the patterns, where the verbs are likely to be used to denote mental processes, check occurrences of the patterns in ‘Alice in Wonderland’ and compare them to their translations to Russian. The important point about this methodology is that it allows us to study more frequent instantiations of the conduit metaphor (or similar mental phenomena) in a corpus that lacks semantic annotations (as it is frequently the case).

As for carry, it is relatively rare, 7 instances in the story, among which it is easy to find the only reference to the conduit metaphor:

(3.8) while more and more faintly came, carried on the breeze that followed them, the melancholy words...

However, the study of collocations of carry in the BNC gives more examples of patterns that are mostly used in this function: carry a headline, a message, a report, a risk, a warning, as well as carry unanimously (a decision after voting).

Get is much more frequent, it belongs to the first 50 most frequent words in English and has 107 instances in “Alice” (and 213,169 in the BNC). So it is better to start with the study of its patterns first. Table 3.4 suggests the most important t-score collocates for the immediate right neighbours of get in the BNC (the second column in the table lists the frequency of a bigram, third and fourth list the frequency of individual lexical items). The patterns get to, get through and get there in the list look promising for further investigation. We can also study the nouns that frequently follow get.

Get to reveals the patterns of get to the point, get to the bottom (typically in the sense ‘to understand every detail’), get to the stage, end. Each of the patterns is exemplified by 20–100 concordance lines in the BNC, so for tetragrams they are very frequent. Get through and get there do not produce statistically significant patterns related to communication, whereas the study of nominal groups (get []\{0,2\} [pos="NN."]) shows that get is very frequently used in this function. Nouns that frequently follow get include impression, message, feeling, idea, answer (often, right or wrong), information, opportunity, etc (other frequent collocates are money, job, chance, advantage, key).

All the patterns are associated with a reinterpretation of the physical process of transmission in terms of communication. In the case studies in
<table>
<thead>
<tr>
<th>Word pair</th>
<th>Joint frq</th>
<th>Freqc 1</th>
<th>Freqc 2</th>
<th>T score</th>
</tr>
</thead>
<tbody>
<tr>
<td>get a</td>
<td>21976</td>
<td>213689</td>
<td>2139625</td>
<td>117.09</td>
</tr>
<tr>
<td>get to</td>
<td>16113</td>
<td>213689</td>
<td>2498679</td>
<td>84.45</td>
</tr>
<tr>
<td>get out</td>
<td>5011</td>
<td>213689</td>
<td>149892</td>
<td>66.22</td>
</tr>
<tr>
<td>get ta</td>
<td>3632</td>
<td>213689</td>
<td>4713</td>
<td>60.10</td>
</tr>
<tr>
<td>get it</td>
<td>7312</td>
<td>213689</td>
<td>1062204</td>
<td>58.70</td>
</tr>
<tr>
<td>get up</td>
<td>3887</td>
<td>213689</td>
<td>181181</td>
<td>56.07</td>
</tr>
<tr>
<td>get into</td>
<td>3535</td>
<td>213689</td>
<td>157633</td>
<td>53.73</td>
</tr>
<tr>
<td>get back</td>
<td>3238</td>
<td>213689</td>
<td>101717</td>
<td>53.05</td>
</tr>
<tr>
<td>get some</td>
<td>3248</td>
<td>213689</td>
<td>167318</td>
<td>50.65</td>
</tr>
<tr>
<td>get on</td>
<td>4822</td>
<td>213689</td>
<td>705835</td>
<td>47.50</td>
</tr>
<tr>
<td>get away</td>
<td>2282</td>
<td>213689</td>
<td>35715</td>
<td>46.16</td>
</tr>
<tr>
<td>get rid</td>
<td>2068</td>
<td>213689</td>
<td>2604</td>
<td>45.35</td>
</tr>
<tr>
<td>get your</td>
<td>2321</td>
<td>213689</td>
<td>134384</td>
<td>42.16</td>
</tr>
<tr>
<td>get any</td>
<td>2229</td>
<td>213689</td>
<td>121477</td>
<td>41.66</td>
</tr>
<tr>
<td>get one</td>
<td>2471</td>
<td>213689</td>
<td>296118</td>
<td>36.85</td>
</tr>
<tr>
<td>get off</td>
<td>1446</td>
<td>213689</td>
<td>67530</td>
<td>34.19</td>
</tr>
<tr>
<td>get through</td>
<td>1415</td>
<td>213689</td>
<td>81188</td>
<td>32.96</td>
</tr>
<tr>
<td>get married</td>
<td>1072</td>
<td>213689</td>
<td>7796</td>
<td>32.23</td>
</tr>
<tr>
<td>get used</td>
<td>902</td>
<td>213689</td>
<td>19644</td>
<td>28.62</td>
</tr>
<tr>
<td>get home</td>
<td>947</td>
<td>213689</td>
<td>56658</td>
<td>26.80</td>
</tr>
<tr>
<td>get there</td>
<td>1815</td>
<td>213689</td>
<td>318095</td>
<td>26.49</td>
</tr>
</tbody>
</table>

Table 3.4: T-score values for the collocates of *get*
the book we will refer to this, following the terminology used for describing grammatical metaphors (Halliday and Matthiessen, 1999:227ff), as congruent vs. non-congruent expressions.

Then, we can search the patterns in the original English text of ‘Alice’. Let’s consider some examples:

(3.9) *I shall never get to twenty at that rate!*

(3.10) *till she got to the part about her repeating ‘YOU ARE OLD, FATHER WILLIAM,’*

(3.11) *Everything’s got a moral, if only you can find it*

There is also an example that plays on the both source and target domains of the conduit metaphor:

(3.12) *you had got to the fifth bend, I think?*

In saying this, Alice simultaneously means ‘to tell a story up to a specific topic’ and ‘to travel to a specific point’ (as she imagines that the Mouse tells the tale along its tail).

Grady (1998) provides the following set of metaphoric mappings implicit in the “Conduit Metaphor”:

1. Constituents are Contents (words and meanings are contained within sentences and the discourse in general);
2. Achieving a Purpose is Acquiring a Desired Object (a message is received, and its meaning is accessed as a desired object);
3. Information is Contents (meanings are contained in forms);
4. Transmission of Energy is Transfer (a physical transfer is involved in a communicative act);
5. Mental Phenomena are Possessions (the contents of mental phenomena are possessed by a person that acquired them).

The fourth step is the least acceptable treatment in Grady’s analysis, since energy and its transfer are highly metaphoric notions themselves. His analysis uses energy instead of physical movement in order to take into account transmission of acoustic signals, radio waves or electric signal in Internet, which involve no perceptible motion, but only energy transfer. However,
the physical account of energy transfer is not grounded in the phenomenal experience: a person interprets the result of message acquisition, while the process, through which the message is acquired, is considered as an (instantaneous) movement in the physical space.

Also the fourth mapping is not supported by examples in the corpus, so only mappings 1-3 and 5 are considered below.

On the other hand, examples (3.9)-(3.12) call for an additional mapping:

6. Communicating is Travelling (a sequence of containers is involved in communication, so they constitute a train of thoughts or words).

This mapping means that the author refers to a story as a sequence of physical locations, so that story telling is represented by a trajectory: at any moment there is a point at which the speaker has just arrived. Also, the author may refer to already expressed sentences and to what is going to be expressed, as points along the trajectory. In linguistic terms, this can be expressed by a reference to landmarks of the story telling space, cf. examples (3.9), (3.10) and (3.12).

The “Conduit Metaphor” pervades various European linguistic cultures, according to Whorf’s concept of the Standard Average European (SAE) mould. All the examples (3.9)-(3.12) are translated into Russian also using the “Conduit Metaphor”. However, the set of metaphoric mappings used in a particular situation conforms to the lexicogrammatical possibilities of a particular language.  

(3.13) For (3.8)

D: veterek s morja donosil grustnyj napev
the breeze from the sea brought a melancholy melody

N: i vsjo tishe i tishe zvuchali gde-to pozadi obryoki unylogo pripeva
snatches of a dismal song were more and more faintly heard somewhere behind

(3.14) For (3.9):

D: tak ja do dvatsati nikogda ne dojdu!
in this way I’ll never come to twenty

N: ja nikogda ne doberus’ do dvatsati!
I’ll never reach twenty

---

9 Translations with insignificant differences are omitted. D stands for Demurova, N for Nabokov, Z for Zakhoder.
3.5. THE CONDUIT METAPHORS IN TRANSLATIONS

(3.15) For (3.10)
D: pokonanedoshldapstoircshisinejgusenitsej
until she came to the meeting with the Blue Caterpillar
Z: itol’kokogdasonadobralasjdetojomesta,gdechitalaSinemu
Chervjakustihi
and only when she got to the part, in which she read verses to the
Blue Worm,

(3.16) For (3.11):
D: Vovsjomest’svoja moral’,nužhnotol’kounet’ejonajti!
Everything contains its moral, if one is able to find it
N: Uvsjakovveschiest’svojamoral’–tol’konuzhnoejonajti.
Every object has its moral, one should only look for it

(3.17) For (3.12):
N: Vykazhetsjadoshlipyatopogiba?
it seems that you came to the fifth bend
Z: Vyonostanovils’napijatopovorote?
you stopped at the fifth turn

Almost all examples interpret communication as transmission of physical objects, with the exception of Nabokov’s interpretation in (3.13), where neither communication nor transmission is implied, the sentence describes the perception of a sound, as a mental event. In (3.15), Zakhoder introduces the “Constituents are Contents” mapping, since the translator ‘opens’ the container for the event “Alice reads verses”. The “Achieving a Purpose is Acquiring a Desired Object” mapping of the original text is kept by all translations of (3.11). However, the “Mental Phenomena are Possessions” mapping is reconsidered in Demurova’s translation as “Constituents are Contents”. In (3.12), the “Communicating is Traveling” mapping is even amplified in Russian translations, which use explicit motion words (come and stop) instead of the generic get in English.
Chapter 4

The lexical model

4.1 Introduction

The analysis of the logic-centred approach to lexical meanings presented in Sections 2.2.1 and 2.2.2 is mostly negative. It confirms that there is a mismatch between concepts defined as senses in a dictionary and uses of words treated as instances of such concepts. What are implications of these problems for modelling uses of words? After all, any dictionary is finite (in practice) in comparison to the infinite amount of situations in which words can be used. How can we describe lexical items in terms of their uses without resorting to the theoretically abominable proposition that each use is unique and is selected by the speaker at just this moment for just this purpose?

The case studies presented in Sections 2.4 and 3.5 are positive. They describe uses of words from the viewpoint of their function in the ongoing communication. However, they do not reach the precision of analysis of individual words, as offered in the logic-centred approach. In particular, the dictionary metaphor is computationally tractable: it is based on a mapping between types (concepts) and their instances, as Figure 2.2 based on WordNet suggests. The open question is whether there is a computational mechanism for describing meanings as uses, so that it can be applied to tasks oriented to natural language processing, but the one that escapes the enumeration of all possible uses of a word as the definition of its meaning.

In his “Introduction to Functional Grammar” (IFG) Halliday refers to two possibilities for describing the grammar within the systemic-functional approach: as systemic and as functional (Halliday, 1985:xv). A systemic analysis assumes that language is a resource for communication and proceeds in terms of interrelated options for expressing meanings. The choice of an
option entails the possibility of some other options, closes the possibility of others, and applies certain realisation statements. A functional analysis starts with the same assumption (language is a resource for communication), but proceeds in terms of metafunctions and shows the functional structure of linguistic constituents (typically groups or clauses) that realise particular options from the systemic descriptions.

The systemic model captures the relationship between various options possible within the communicative resource offered by language, whereas the functional model can be used as a tool for the analysis of texts (both theoretical and practical varieties of analysis). In IFG Halliday focuses on the functional aspect of the description, for instance, he describes the structure of the finite Clause in terms of types of Processes (e.g. it can have a mental process, like see), and Participants (e.g. the Senser and the Phenomenon, in the case of see). Functional descriptions are typically better suited for introductory purposes, because they start with less abstract categories that are easy to recognise in actual instances of text production. This is the reason why the introductory description of lexical semantics in Section 2.4 is also functional and not systemic. That study considers the contribution of individual terms, such as good into the ongoing discourse. However, the complete model requires an explicit statement of the relationship between meaning intentions and their realisations. This is also the requirement for a computational model to be used in natural language understanding, generation or translation.

The goal can be achieved using the mechanism of the systemic network, which allows both a theoretically-sound description of language as a resource and computationally tractable. The systemic network is a network of interrelated choices, semantic conditions underlying the choices and realisation statements following the choices. System is used in systemic linguistics as a technical term referring to each single set of choices, while feature is another technical term referring to choices themselves. For instance, the classification of the English mood starts with a system consisting of two features [indicative] vs. [imperative]. Semantically, it corresponds to the opposition of speech acts referring to exchange of information vs. issuing commands. The choice of [indicative] leads to another system that includes another two choices [declarative] and [interrogative], which are realised syntactically by the respective order of two syntactic functions: Subject and Finite. Further on, the choice of [interrogative] entails the choice between two types of questions: [yes-no] and [wh-interrogative]. In syntagmatic terms, the first choice orders the Finite before the Subject, whereas the second choice adds a wh-element which is added in the beginning of the
4.1. INTRODUCTION

sentence. In technical terms the relationship between [indicative] and [interrogative], as well as between [interrogative] and [wh-interrogative] is described as delicacy. For instance, the relative order of the Subject and the Finite in the case of choosing [wh-interrogative] is said to be defined in more delicate systems. In addition to the classification of mood we can describe the moves of interlocutors in a dialogue, for instance, as [initiate] or [respond]. In the end this gives us the network shown in Figure 4.1.

Each move is described as a parallel choice of two systems (this is graphically represented by the opening curly bracket). Realisation statements that follow the choice of features are given in boxes below respective features. The most frequent realisation statements constrain the order of constituents (on the diagram this is denoted by ^), the presence of a constituent (denoted by +), classification constraints (denoted by a semicolon, such as Finite:Aux, i.e. Finite is realised by an auxiliary verb), the realisation by a specific lexical item (denoted by !, e.g. Subject!it), or conflation (Wh/Subject, i.e. that the wh-element and the subject are realised by the same constituent, for instance *Who knows the number of my room?*). As a matter of convention, when features from the systemic network are discussed in text, they are presented as enclosed in square brackets, e.g. [wh-interrogative]. For more information on representing lexicogrammatical resources in these terms, see (Halliday and Matthiessen, 1999:39ff).

Figure 4.1: Demo of a systemic network for mood in English
4.2 Lexis in systemic linguistics

However, networks drawn in (Halliday and Matthiessen, 1999) mostly concern grammatical features. The same is true for the work on the NIGEL lexicogrammar (Matthiessen, 1996) and many other studies. In the beginning of IFG, Halliday writes:

> in order to make explicit the fact that syntax and vocabulary are part of the same level in the code, it is useful to refer to it comprehensively as ‘lexicogrammar’, but it becomes cumbersome to use this term all the time, and the shorter term [grammar] usually suffices. (Halliday, 1985:xiv)

After this remark he continues with exploring topics related to the grammar in more narrow sense and pays much less attention to lexical items. Of course, a comprehensive study of the English grammar cannot completely exclude topics pertaining to lexical semantics, so in (Halliday, 1985) they are addressed to some extent in chapters on types of processes and cohesion.

However, the distinction between the lexicon and the grammar was clear from the outset of systemic descriptions. According to (Halliday, 1961), word is a grammatical category within the ranks of clause, group, word, morpheme. A word is defined in terms of its contribution to the meaning of a group and integrates contributions of morphemes. On the contrary, lexical item is a lexical category and can appear at any rank of the grammatical description. For example, the following expressions can be considered as lexical items: dialog box (a nominal group), How are you (clause), -able¹ (morpheme).

When a lexical item, e.g. a nominal group, is realised in a grammatical environment, its elements carry syntactic functions, for example, dialog is Classifier in dialog box. In English, the difference between box and dialog box may not be evident at the level of lexicogrammatical realisation, so the lexical item dialog box can be represented in the grammatical structure by a single unit, i.e. in technical terms a lexical item is represented as a word with a space inside it. However, in Russian the modifying adjective agrees in the case and number with the noun head, e.g. dialogovoe okno, dialogovogo okna, dialogovym oknom. This requires that both grammatical structures and lexical meanings should be represented at different, but intersecting levels of description, but this leaves open the question of principles for describing lexical items in SFL.

¹It is considered in the example by (Lyne, 1988): Do you think that the bike is pushable?
One comprehensive account of the English lexicon, which is functional and cognate to systemic linguistics, is the Collins COBUILD English Dictionary, see (Sinclair, 1987), and the pattern grammar approach (Hunston and Francis, 2000) which followed from it. However, descriptions in a human-oriented dictionary lack formal mechanisms for dealing with meanings of lexical items in a computationally tractable way, i.e. such descriptions cannot be directly used in computational applications for language understanding and generation or machine translation. Chris Butler (1985:128–137) overviews the topic of the relationship between lexical and grammatical knowledge in systemic descriptions, but he also does not consider the use of systemic networks.

Some other scholars have described the behavior of lexical items using systemic networks as the formal mechanism. This includes the study of lexis as most delicate grammar (Hasan, 1987), the study of lexical options for adjectives within the lexicogrammatical network (Tucker, 1998) and the polysystemic model, treating lexis as a separate stratum (Wanner, 1997).

The approaches address some properties of the behaviour of lexical items, but remain restricted in certain respects.

The first approach is defined by the slogan “Lexis as most delicate grammar”, which assumes that lexical choices are the most specific choices in the lexicogrammar and lexical features are subsumed under grammatical features, e.g. all lexical possibilities for reporting a speech are subsumed under the general class of processes of saying, see, for example, (Halliday, 1961), (Hasan, 1987). The second approach is the polysystemic model, which assumes that the lexis and the grammar are two separate strata, which are
based on different principles of description and different systemic networks, cf. (Lyne, 1988), (Wanner, 1997). This approach follows Firth’s work (Firth, 1957) on collocations. Another computational instantiation of the same approach is the MOOSE lexicalisation mechanism by (Stede, 1996). The separate lexical stratum provides resources for converting situation specifications into lexicalised semantic specifications consumed by the grammar. For instance, the situation specification BATH(X), where X is the actor, can be converted to two lexicalised specifications, one with the non-directed process *bath*, and one with the directed process *take* and the goal *bath*, each of which can be used by the grammar to produce *X bathed* vs. *X took a bath*. However, the separation of the lexicon and the grammar can lead to reduplication of grammar-induced properties in the lexical network and vice versa. Tucker’s model is between the two extremes in Figure 4.2, since it does not posit a specific order of delicacy for the grammar and the lexicon like (Hasan, 1987) and does not define a separate stratum like (Wanner, 1997).

However, one criticism is applicable to these three approaches at once: unlike COBUILD, they are not functional enough, in the sense that they do not relate lexical items to communicative functions underlying their uses. Also (Wanner, 1997) describes general principles of his model, but does not provide a detailed study of a set of lexical items in terms of the model. (Hasan, 1987) treats few lexical items, like *strew*, *spill*, etc, which are also quite rarely used. (Tucker, 1998) covers the complete domain of English adjectives, but treatment of specific classes of adjectives remains cursory, for instance, non-dimensional senses of size adjectives, like *a broad coalition*, *a little girl*, or *high interest rates* are completely missed.

### 4.3 The three views on lexical resources

The current study mostly concentrates on purely linguistic aspects, more specifically lexical semantics from the viewpoint of communication, leaving aside psycholinguistic and sociolinguistic aspects of communication. In the case of lexical semantics, this topic corresponds to studying the *potential* for communication provided by lexical items and *instantiation* of the potential, when lexical items are used in context. Thus, the approach is, in a sense, generative: the finite set of lexicogrammatical resources is used to communicate a large set of speaker’s meaning intentions in an infinite amount of contexts.

Let’s start with considering the model proposed in the previous chapter
4.3. THE THREE VIEWS ON LEXICAL RESOURCES

in terms of the systemic network. The essence of the proposed approach
to describing meanings of lexical items from a communication-centred per-
spective was presented in Figure 2.3 in Section 2.3. It assumes two par-
ties involved in communication: the speaker with communicative intentions
and the hearer. The speaker uses lexicogrammatical resources for express-
ing communicative intentions according to the context of situation. The
meaning intentions are realised as sequences of physical signals (printed
characters, acoustic waves, non-verbal gestures), which are interpreted by
the hearer. Interpretation is based on the set of linguistic resources and
the context of situation, which in the case of successful communication are
largely shared with the speaker. The description of meanings of lexical items
in terms of their uses consists in defining the potential of possibilities for re-
alising meaning intentions by means of lexical items and in describing the instantiation of the potential, when lexical items are used in context. Thus,
the approach is, in a sense, generative: the finite set of lexicogrammatical
resources is used to communicate a large set of speaker’s meaning intentions
in an infinite amount of contexts.

Halliday and Matthiessen (1999:504) offer the following trinocular view:

1. “from around” the lexicogrammar: to consider what choices are avail-
able in the lexicogrammar, for instance, [declarative] and [interroga-
tive] within the system of indicative clauses;

2. “from above” the lexicogrammar: to look at the distinction in meaning
which is reflected in the choice between [declarative] and [interrogative]
in the system;

3. “from below” the lexicogrammar: to consider differences between [declar-
itive] and [interrogative] in terms of their syntagmatic realisation.

In the case of lexical semantics, the view “from above” means a description
based on communicative intentions; the view “from below” is based on
individual lexical items, while the view “from around” describes choices
available for uses of lexical items in a lexical domain (verbs of motion, size
adjectives, words denoting emotions, etc).

The study of uses of lexical items allows the same set of views. The view
“from above” means a description based on communicative intentions, e.g.
similar to information found in thesauri; the view “from below” is based on

---

Footnote: The notion of intention used here refers to any piece of semantic information that is
going to be expressed by the speaker, thus including not only conscious intentions, but
also such distinctions as those between types of carriers in riding, flying, swimming, etc.
individual lexical items, so it is similar to information found in dictionaries, while the view “from around” describes choices available for uses of lexical items in a lexical domain (verbs of motion, size adjectives, words denoting emotions, etc). The three views can be considered in greater detail.

4.3.1 The view “from around” the lexicogrammar

This view is based on the systemic network, which includes paradigmatic classifications of features and syntagmatic realisations, which specify consequences in the choice of features. In the case of lexical semantics, realisation statements constrain the choice of lexical items from the lexical stock available in language (by means of classify and lexify statements). The potential of uses is represented by choices of features in the systemic network.

According to the same principles as used for development of systemic grammars (Martin, 1987), a feature in the lexical network is motivated, if it has some reflex in form, i.e. if several lexical items are available for expressing a group of communicative intentions, they correspond to several distinct features, e.g. go, drive, fly for distinguishing between vehicles, or if there are grammatical constraints on realisation following the choice of a feature, e.g. the source or the destination is obligatory. In principle, each type of use should correspond to a unique set of features, which should be related to respective communicative intentions. For instance, escape, move out, quit, run away can refer to leaving a place without an intention to return back. However, the verbs provide different lexical realisations for different communicative intentions, respectively, leaving a place of confinement (escape), relocating to a new home (move out), leaving a place under external pressure (quit), escaping from home secretly (run away). Such properties of the place or the cause of leaving are reflected by features of the systemic network, which end up with selection of a lexical item best suited for expressing the intention.

The lexical network also addresses the issue of differences between uses of words in different languages. The systemic research on the grammar of typologically related and unrelated languages shows that least delicate, more general, choices tend to be shared across languages, while more delicate choices tend to be language-specific (Bateman et al., 2000). The same principles guide the description of lexical meanings by means of the systemic network. For instance, many languages have the type of motion verbs and also distinguish between the motion towards or away from a reference object, e.g. enter vs. leave, eingehen vs. ausgehen, vijti vs. vyjti. Moreover, the three languages have specific options to address the fact that a person leaves
4.3. THE THREE VIEWS ON LEXICAL RESOURCES

a vehicle at the end of the journey, e.g. *get out, not go out or leave*, but further options are language-specific: English distinguishes between public and private transportation means (*get off a train, bus, plane vs. get out of a car or boat*), German uses *aussteigen* for all types of vehicles, while Russian distinguishes between the relative height of the vehicle: *vyjti* for any vehicle, but *sojiti* for leaving a vehicle that is higher than the ground (mostly planes and ships, but also, less frequently, trains). Thus, language-specific systems and features should be added to the lexical network, which represents both commonalities and differences between languages.

4.3.2 The view “from above” the lexicogrammar

This view is based on information that is available in the context of situation and guides the options in the lexicogrammar. In this view, features in the network are themselves used for realisation of higher-order options related to communicative intentions. For instance, if we consider the uses of *come* vs. *go*, we can encode them in the systemic network as the opposition of *towards* vs. *away-from*, which are realised by a multitude of verbs, including *come* and *go*. However, when we consider principles that guide their choice, then the choice follows such principles as the presence or absence of the speaker at the destination point (Fillmore, 1983). In the case of a computational model, sources of information can be implemented by various semantic experts, which relate options available in the systemic network to properties of the situation. For instance, Fillmore (1983) considers two rules for distinguishing

\[(4.1) \text{ We’ll come there at dawn.} \]
\[(4.2) \text{ We’ll go there at dawn.} \]
\[(4.3) \text{ Let’s go there at dawn.} \]
\[(4.4) * \text{Let’s come there at dawn.} \]

The first rule implies that *go* is possible, when the destination is distinct from the speaker’s location at the speaking time. The second rule implies that in the *come* is possible, when either the speaker or the addressee is at the destination point at either speaking time or reference time. In (4.1), unlike (4.2), the addressee is excluded from the group designated by *we*, because *come* is possible only, if either the speaker or the addressee is there at dawn. The version that allows the inclusive reading only (with *let’s*)
leaves no possibility to use *come*. This is exactly the type of information described in the upper stratum controlling choices in the lexicogrammar.

Another example may concern the difference in the type of the carrier of motion, e.g. motion by foot, wheels, wings, using a boat, etc, can guide the choice of options in the lexicogrammar that ultimately realise *go, drive, fly, sail*. Note that the choice is not the straightforward mapping of the type of motion to features, it also depends on traditions for expressing certain events and objects and appropriateness of the expression in the current discourse context. For instance, the domain of flight, bus or train schedules in English favours generic terms for designating the departure of a vehicle, like *leave* (*my flight leaves in less than an hour*). In German and Russian, the possibilities for using generic references are more restricted: in Russian the departure of a plane should be referred as flying (*vyletait*), while other types of departure are expressed in a generic way (*vykhoditj*, *leave*), but in German all cases of departure according to a schedule are expressed by verbs specific to the type of a vehicle: *abfahren* (*leave by bus, train and boat*) and *abfliegen* (*leave by plane*).³

³On the few exceptions, see the chapter on verbs of motion below.
4.3. THE THREE VIEWS ON LEXICAL RESOURCES

Now let’s proceed to a more formal account. Figure 4.3 shows a systemic network for the lexicogrammar of smaller age in English. First, there is a choice between the age of a sibling that is younger than the referent (an example from the BNC: *My little brother is in prison*, where his current age is not a question) vs. the absolute age specification, which in its turn leads the choice between the reference to the age of a child (*little*) or adult (*young*). In computational terms the choice itself is made by consulting the knowledge base with information about discourse referents, this is the *Age-Type-Q* inquiry in Figure 4.3 (Matthiessen, 1983).

4.3.3 The view “from below” the lexicogrammar

This view maps individual lexical items to possible combinations of features in the network. The structure of description is similar to distinction between senses available in dictionaries: a set of features often corresponds to a sense. There are two possibilities with respect to the relationship between features in the systemic network and senses in a dictionary. Some features are associated with some lexical items. For instance, *indignant* always refer to an emotion of a specific type, as CCED (1995) puts it:

(4.5) *If you are* **indignant**, *you are shocked and angry, because you think that something is unjust or unfair.*

This assumes that a set of features corresponds directly to a lexical item, in this case, we can assume [anger], [intense-emotion], [directed-at], [existing-cause]. Other sets of features can correspond to a specific sense of lexical item, as detected in a dictionary or it can be more fine-grained than a dictionary definition allows. For instance, such words as *leave, abfliegen* and *vyletat’* can be used to refer to physical motion, specifically to “away-from” motion, or to abstract motion, as in *to leave school*, but unlike *abfliegen* and *vyletat’*, the English verb is not specific with respect to the type of motion and specific meanings of *leave*, as in ‘drive-away’ or ‘fly-away’ are not detected as senses in a dictionary, but they correspond to specific combination of features.

Another difference between features and senses detected in dictionaries, is that the latter are typically based on concepts, while features in the network correspond to functions for using lexical items. On the other hand, CCED (1995) is an example of a dictionary which definitions are oriented towards possible uses of words rather than definitions of concepts, even though its lexical entries are not formal enough to be suited for computational applications and the requirements imposed by the ECD model. Definitions
in CCED (1995) are aimed at human readers, most specifically, learners of English. The dictionary tries to teach them how to use English lexical items correctly by suggesting situations, in which a learner can use a word:

\[(4.6) \text{If you leave someone to do something, you go away from them so that they do it on their own}\]

The ECD uses another concept of correctness: it is aimed at the theoretically correct description of lexical items independently from purposes they are used for. As the result of different intentions, definitions differ significantly. Definitions in COBUILD are informal and circular, for instance, its entries for *go away* and *leave* refer to one another, e.g. compare (4.6) to “If you *go away*, you leave a place and spend a period of time somewhere else, especially as a holiday.” This circularity is not allowed in the formal model of ECD, but it may be of help for human readers of CCED (1995), who, most probably, know the basic meaning of *go away* and *leave* and need to know constraints on their uses in specific situations. However, CCED (1995) lacks formal means, which allow to use the descriptions in computational applications.

On the other hand, there is also some similarity between ECD and systemic descriptions, namely the attention to collocations. In addition to the definition zone, a lexical entry in the ECD has the government pattern zone, which specifies semantic actants of a lexical item, and the zone of lexical functions for describing various co-occurrence relations. Mel’čuk defines lexical functions as functions (in the mathematical sense) that associate keywords with their values (Mel’čuk, 1996). The keyword of a lexical function is a lexical expression, and its value is another lexical expression. Mel’čuk (1996) discusses the set of about 60 standard simple lexical functions that conform to several conditions the most important of which is the availability of an abstract meaning applicable to a relatively large number of arguments to yield a large number of different lexical expressions as their values. For instance, the meaning ‘a year in which February has 29 days’ constitutes a lexical function, because it captures the idiosyncratic relationship between the lexical item corresponding to *year* and the respective terms in various languages, as in *leap year, année bissextile, Schaltjahr, viskosnyj god*, etc. However, it is a non-standard lexical function, because it can be applied only to words expressing the meaning of ‘year’.

One of the standard lexical functions is *Magn*, the lexical function for expressing the intense degree. For instance:

\[(4.7) \text{Magn(to condemn)}=\text{strongly}\]
4.3. THE THREE VIEWS ON LEXICAL RESOURCES

\[
\begin{align*}
\text{Magn}(\text{patience}) &= \text{infinite} \\
\text{Magn}(\text{range}) &= \text{wide} \\
\text{AntiMagn}(\text{range}) &= \text{narrow}^4 \\
\end{align*}
\]

In terms of systemic linguistics, the government pattern zone of a headword corresponds to the definition of participants in the situation defined by the headword and to constraints on their morphosyntactic realisations. For instance, the government patterns for *to send* are $X$ sends $Y$ to $Z$ and $X$ sends $Z$ $Y$, where $X$, the sender, is a human being, $Y$, an object, is in the oblique case, while $Z$, the addressee, is another human being, which in the second pattern is preferably realised by a pronoun or a proper noun (a full scale nominal group is unlikely here). In the systemic network the restrictions on the participants can be modelled using the *classify* realisation statement.

Note that the systemic notation can do more than the government pattern does. The government pattern from the ECD model *defines* the morphosyntactic properties of participants, whereas the realisation statements from the systemic model *constrain* them. This means that some realisation statements can be added in less delicate systems to achieve generalisation, as the choice of the accusative case for realisation of direct complements in Russian, while others can underspecify the morphosyntactic realisation assuming that the exact form will appear as the result of the application of constraints from other systems.

The notion of lexical functions can be also profitably used in systemic studies as a method for describing the contribution of lexical expressions. For instance, in

\[(4.8) \text{Factory shops contain a wide range of cheap furnishings,} \quad \text{D E:Magn T}\]

\text{Magn} is a functional label that refers to the size of a collection of furnishings (D, E and T stand for Deictic, Epithet and Thing respectively). Another lexical function used by Mel’čuk is \text{Bon}, a positive evaluation of an object, e.g.

\[(4.9) \quad \text{Bon(advice)} = \text{sound} \\
\text{Bon(aid)} = \text{valuable} \\
\text{Bon(description)} = \text{thorough} \\
\text{Bon(knowledge)} = \text{extensive}\]

\(^4\text{Anti} \text{ is another standard lexical function, which means an antonym for its keyword. As we see here, some lexical functions can be combined.}\)
CHAPTER 4. THE LEXICAL MODEL

Another class of lexical functions is $\text{Oper}_i$, which defines a verb which actor is the i-th participant of the situation introduced by its argument, and which actee is the argument itself. For instance, if the situation is described as ‘resistance of X to Y’, then

\begin{align*}
(4.10) \quad \text{Oper}_1(\text{resistance}) &= \text{offer} \\
\text{Oper}_2(\text{resistance}) &= \text{meet with}
\end{align*}

The notion of lexical functions is similar to the notion of a collocation, which is widely used in systemic linguistics, as it was introduced by Firth (1957), who defined them as “the company a word keeps”. For instance, offer and meet are also the most important verbal collocates of the word resistance in the Bank of English. At the same time the name of a lexical function explicitly defines the relationship between the two words, while a collocation simply states that there is a relationship. So the mechanism of lexical functions is very useful. What is good is that it can be modelled in the systemic network even in two ways. The easiest way is to use the direct lexicalisation statement, e.g. Quality!wide in the case, when the lexical item range is used with the functional role $\text{Magn}$. Another method can use a realisation statement called preselection, which can be thought of as a specific subtype of class restrictions that causes the choice of other features in the network (Halliday and Matthiessen, 1999:43–44).

However, the systemic network can once again do more than lexical functions do. A lexical function offers the word-to-word mapping, because it simply links two lexical expressions: its output is “all or nothing”, i.e. it finds a word corresponding to a lexical function (if there is any) and outputs a value (another word) without any consideration of possible communicative intentions of the speaker or the context. For instance, the result of an application of $\text{Bon}$ to advice is always sound, even though an advice can be good, best possible, helpful, independent, practical, professional, etc. At the same time, the lexical choice in the systemic network is controlled by combination of several features, for example, including semantic specification of the situation (the consideration from the ideation metafunction) or rhetorical properties of the interaction between the speaker and the hearer (the interpersonal metafunction) or the ongoing discourse (the textual metafunction). The system of such choices for choosing the value of $\text{Bon}$ for advice can be related to a more general system of APRAISAL (Martin, 2000).
4.4 Case Study: the description of “to burn”

4.4.1 The lexical domain

The communication-centred answer to problems of the logic-centred approach discussed in Section 2.2 is based on the fact that language offers a system of resources for expressing specific kinds of situations. For instance, *to burn* can be used to refer to many different things, or, since it is a verb, can be used in various contexts. At the same time, there are many other words in the semantic domain of burning, for instance, in English the list includes, at least, *glow, flare, gleam, blaze, glare, be on fire, set fire to,* etc. In German and Russian the situation is even more complicated because of the large number of words that can be derived using prefixes and can refer to various aspects rendered by a single verb in English, e.g. in German for *brennen* (the direct translation equivalent of *to burn*) we also have *abbrennen* (the process of burning), *anbrennen* (to start burning, but also to scorch), *ausbrennen* (the result, as in *to burn out*), *einbrennen* (to leave marks by burning), *verbrennen* (also the result, but in other contexts, as in *to burn a finger*), etc.

The very nature of the systemic approach, which treat language as the system of resources available for communicating and understanding, calls for the study of a range of means available in a particular lexical domain. The second part of the book offers large-scale studies in three lexical domains (size adjectives, verbs of motion and words referring to emotions). At the same time for the purposes of an introductory study in the proposed framework we need to consider a small-scale network covering uses of very few words from the domain. So the study in this section is restricted to *burn* and *glow* in English. As for Russian the description covers *goretj* and *svetitj* as well as their most important aspectual counterparts *sgoretj* and *svetitjsja*, because otherwise the study will be utterly incomplete.

Also, the number of possible types of uses is potentially infinite. The study of the behaviour of a word utilises descriptions available in dictionaries, however, there are often gaps in the description of dictionaries (see examples in Section 2.2). What is more, the design of lexical entries in traditional dictionaries corresponds to the goal of their production: to help their users in distinguishing between separate senses. So not all dictionaries are good in explaining the context and purposes of basic word uses. This is especially true of formal dictionaries, such as WordNet, or ethymologically oriented dictionaries, such as the Oxford English Dictionary. Out of this reason the communication-oriented study like this is based on a study of
concordance lines produced from a representative corpus. For expository purposes, the study is once again restricted in another respect. It uses a small concordance of about 20–25 lines for each word, see Figures 4.4 and 4.5. English lines were taken from the BNC, Russian ones from the Russian Reference Corpus; the lines were sorted according to the immediate right neighbour. In addition I will consider some uses of burn and goretj discussed above in Section 2.2.

Dictionaries or larger concordance studies can show a much wider range of possible uses for the words in question, but even the selected lines show some interesting results. On the other hand, the limit of the study gives the possibility to pay attention to every concordance line in the description. The study begins with the same approach to the analysis of lexis in texts as used by John Sinclair (1988), namely that the totality of concordance lines for a lexical item shows regular patterns of its uses. The task of the study is to define the set of general sense groups to adequately classify each and every concordance line.

However, for the sake of space, the types of uses of the English verbs burn and glow will not be considered here in detail. It is clear that line 21 in the concordance for burn and lines 2, 21-25 for glow express an emotional state, whereas lines 1, 4, 13, 14 and 17 refer to deliberate destruction of physical objects, such as leaves, houses, lists, etc., as the effect of fire (line 13 is an apparent joke referring to the case, when consecrated wine spills over the credence table). The difference between the contexts of uses of burn with emotion and glow with emotion can be observed only in a larger concordance. The study of the lines output by the queries burn with and glow with show that they collocate with different emotion types. Typically burn with occurs in the context of anger, desire, pain, rage, shame, whereas glow with collocates with pride, warmth, desire, happiness, pleasure, satisfaction, as well as health.\footnote{CCED (1995) treats the latter as a separate sense of glow by associating it with skin: \textit{If someone’s skin glows, it looks pink, because they are healthy}, but as concordance line 22 shows skin is not the necessary collocate.}

So burn with collocates with negative or at most neutral things, such as curiosity or desire, while glow with is clearly positive. Even when the two verbs collocate with the same noun, as it is the case with desire in lines 21 in the both concordances, the context of Jay’s body glowing with desire is definitely positive, as the desire is clearly approved by the author, whereas the context of a sex that burns with desire shows the disapproval of the same type of desire by another author in a different text.
4.4. CASE STUDY: THE DESCRIPTION OF “TO BURN”

**burn** (rank: 1829, frequency: 50.91 ipm)

- it be like ” Forty-six again ? Houses burnt – gardens and orchards uprooted –
- and burns easily, or has a tendency to burn – use a Sun Protection Factor greater
- controlling apple scab is to collect and burn all fallen leaves from infected trees
- and Jan Hus preached before being burnt for heresy in 1413; he pays respect
- g and in too short a time had effectively burnt himself out. His attitude to life
- ire since the last war. Both Coleshill (burnt in 1953) and Dunsland (1967) were
- oltage, the compressor in a refrigerator burns itself out. So there must be a volt
- rderer is discovered at his trial to have burnt himself rescuing two little children
- t, the intense and lurid splendour which burns like gold, and bathes like blood.
- plash from the river, his defiance would burn lower and he would give his name.
- wood, the credence table must be either burnt or eaten. Still; look on the bright
- eous combustion. They should be either burnt or laid out flat to dry, or kept in
- ge dead thing, smouldering perhaps, but burnt out, longer conveying any meaning.
- nce tolerated). It is also traditional to burn spices for the deceased; to cover the
- have half an hour, the first thing is to burn the lists. And the petitions. " " A
- utter to stay in one place long enough to burn the surface of the timber. This gives
- r they work, the more calories they will burn up, and the leaner they will become.
- lal power of nature. The main spars were burnt where the control cables had been me
- fore the ladies. So which sex is it that burns with desire for an encounter that ca

**glow** (rank: 5808, frequency: 9.20 ipm)

- as the way in which newly-lifted potatoes glow " like children's shoulders after a d
- Maggie too began, faintly, pearly, to glow. " You can give it away, magic, "
- ed. Whenever my atoms get excited they glow. They're glowing for you and your
- the hotplate of the stove glowing a faint orange. They began to set
- long fits of massaging, until his scalp glowing a wealed pink and his arms dropped
- him suck his chocolate. His neck began to glow again. He moaned softly. " I want
- hat she had saved money and his life could glow again in the concentration of her att
- ervention. The screens of our word-machines glow as serenely as illuminated manuscri
- gain, hands clapped a rhythm, and faces glowing briefly as pipes were lit. Cameron
- Early Learning Centre climbing frames, glowing cosily out from safely bohemian pin
- its themselves looked asleep, the heather glowing dust-blue in the hazy light, and th
- itional treatments from DeCleor you’ll be glowing from top to toe by the end of the w
- into the darkness, the end of his cigar glowing like a red warning signal as the su
- al glazed windows of the central part and glowing like honey on the yellow walls. Pe
- m the fissure poured lava, blood-red lava glowing like the fires of hell in the night
- ed attitudes. Then, as lanterns start to glow on wooden balconies and red gleams
- ed in the rain, and with every bright flower glowing out through the dismal weather in t
- gold but gold and silver. Hermione was glowing radiant in the dusk, an intense si
- s quickly. Anyone who is not relaxed and glowing with a sense of well-being at the w
- sheets as they slid into bed, Jay's body glowing with desire. And Lucy's silvery war
- " You look wonderful," said John, " glowing with health. Shall I pour you a cu
- It is... inspiring. " " Really? " Erika glowing with pleasure. " Yes, really. Tha
- s and sweater and his weather-beaten face glowing with satisfaction and well-being. L
- was ready. And St Margaret, her cheeks glowing with the realisation of her strengt

Figure 4.4: Concordances for burn and glow in English
CHAPTER 4. THE LEXICAL MODEL

goret' (rank: 874, frequency: 124.72 ipm)
sgoret' (rank: 3702, frequency: 28.02 ipm)

1 cherez neskol'ko let nash teatr vse-taki sgorel, chto lishnij raz podtverzhdaet tu pra
2 heskaja obstanovka? Ne tol'ko tof'janiki gorjat, no i prosto les. A vozmozhnost' lesn
3 omnoe kolichestvo materialov, kotorye ne gorjat, no pri pozhare vydeljajut jadovitve v
4 torogo iskazili vsju istoriju veka; mozg sgorel, obessilel, zhivet xaoetchno, soznav
5 v... Levka odno vremja vseh erudoj gorel. Novyj Levkin batja byl gromadnogo r
6 blagorodnym sochuvstvii ko vsem bla
dom byvshes'. finansovogo direktora ko
7 ka prodavat' ne sobirat's'. V X'justone sgorel
8 go zhira i, pyuznuz na kraeyny ugeol'kax, sgorial golubovatym plamenem. Chiku pokaza
9 roj izvestkoj... V metallicheskih bochkax goretli kostry... Nashu model' berezhno opustil
10 eksja i prodolzhal rabotat', vtoroj glupo sgorel na kostre. Pervyj put' javno razumne
11 y gendirektora v godovuiu ego bankrotstva sgorel ne sluchajno, i namerena rassledovat'
12 u dvora, a potom ona vidit, chto na nem gorit rubaxa, a on ee pytateja pogasit'.
13 govorku - pro togo, na kom lichshe vseho gorit
14. nagolodalis' i aj... na vseh komnate goret
15 ke i sorochke Mu/en... Oksforskih zaponki goretli v luchax poludennogo solnca. Muzheti
16 s', vyinosil svoi stendy, poka ne nachali goret' volosy na golove. Uspej U Lebedeva bo
17 ustanovlennoj norme. Evrosojuz tozhe ne gorit zhelaniem delit's', i vy
18 vat' novuju staruju strukturu. Kiselev ne gorit zhelaniem dogovarivat'sja s "Lukojlo
19 vorov ob aviaperevozki Evropejcy davno gorjat
20 iz nix, chto regulirujut rynok stali, ne gorjat
21 ty, aspiranty - gotovilis' k chetvergu, gorja zhelaniem sotrudnichat' s rossijskimi
22 om provode. Sejchas den', no lampochka gorit zhetym, xudosochnym svetom, kotoryj

svetit'sja (rank: 2876, frequency: 36.98 ipm)

svetit' (rank: 6691, frequency: 13.78 ipm)
1... Publike (to est' nam s vami) nichego ne svetit, a sogrevee nas tol'ko nadezhda
2 od inostranencam. Togda uzh nam nicherne ne svetit', i lichshe sejchas xoroshie den'g
3... Pomnu, vyjesh' noch'ju vo dvor, luna svetit, ogromnaja, belaja. Ten' ot pl
4 most'. I ehe reka, ee ne skroeshi'. Ona svetilas', otrazhaja zvedy, ee izgiby
5 a den'gi (a nekotorye soznatel'no ne xotjat svetit', slovno radujas', chto my edem
6 hudesnyj sentjabr'skij den'. Solnze krotko svetilo, slovno radujas', chto my edem
7 ja. Tut nikakix zarabotkov pravitel'stvu ne svetit', zato bol'shie rasxody na dolgie
8 etv' s shapkoj syrogo - v potemkax on edva svetilsja... Glebov postojal u o
9... i volosy u nee stojali na golove i svetilis'. JA ehto na vsju zhizn' zapom
10... Rozovo-smugloe nebo na zapade e... svetilos' - snega. Glebov postojal u o
11 x stali v Evrosuju - nam takzhe malo chto svetit... Evropejcy bojatsja, chto ix ryn
12 reograf Ivan Fa'deev). Profychnaja scena svetilas' ehektricheskimi lampochkami.
13 aj... Professor uzhe ne molod, no ego glaza svetjatsja junosheskim zadorom, neumnoj
14 a i yshel v sad... Byla olichnaja noch'.... Naprotiv glybami gromozdlis
15 x stali v Evrosuju - nam takzhe malo chto svetit... Evropejcy bojatsja, chto ix ryn
16 reograf Ivan Fa'deev). Profychnaja scena svetilas' ehektricheskimi lampochkami.
17 aj... Professor uzhe ne molod, no ego glaza svetjatsja junosheskim zadorom, neumnoj
18 a i yshel v sad... Byla olichnaja noch'.... Naprotiv glybami gromozdlis
19... sotrudnichat' s rossijskimi
20... Ne vs

Figure 4.5: Concordances for goret' and svetit'sja in Russian
4.4. CASE STUDY: THE DESCRIPTION OF “TO BURN”

The Russian concordance lines also classify into several groups. The basic use of goret’ (like burn) also refers to the destruction of a physical object by fire, these are lines 1–3, 7, 11, 12, 16 as well as 10 (burning as the execution) and 13, where it is used as a part of the following idiom:

(4.11) Na vore shapka gorit
     on thief hat burns
     ‘the hat burns on a thief’, A thief shows the crime in the behaviour

Both goret’ and svetit’ can also refer to the production of light: lines 14, 15, 22 for goret’ and 3, 4, 6, 8–10, 12, 14–17, 20 for svetit’. Both verbs are also used to express emotions: lines 5, 6, 17–21 for goret’ and 13, 19 for svetit’. Here the situation is once again similar to English: svetit’ (to glow) in the reflexive form can be used for a variety of positive emotions and emotional descriptions, e.g. radost’ju (happiness), zadorom (ardour), umom (intelligence), etc, while goret’ (burn) is not applied to them. Most typically it collocates in this sense with zhelanie (desire), or as gorjaschij vzgljad (fired eye).

Another frequent type of uses of svetit’ is expectation, lines 1, 2, 7, 11, 18, for instance:

(4.12) Tut nikakix zarabotkov pravitel’stvu ne svetit.
     here any profits-gen government-dat not light-3sg.
     ‘In this case the government will receive no profits.’

(4.13) Po vsem prognozam imenno im svetit samoe
     For all prognoses exactly they-dat light-3sg most
     znachitelnoe povyshenie.
     significant increase-nom.
     ‘According to all expectations they will see the most significant increase (of prices).’

The use implies a metaphor that the light brings an outcome of an event, which is treated as the grammatical subject (in the nominative or genitive case, because of the negation in 4.12), and the ‘beneficiary’ of the result, which is always negative, is realised in the dative case.

Now we are ready to develop a lexical model in the domain using the systemic network.
4.4.2 *burn* and *glow*: the view from within the network

The basic distinction between the uses concerns the reference to a physical process or its metaphorical interpretation which has nothing to do with the production of light or heat. This is valid for English and Russian, for instance, cf. lines 20–25 for *glow* to lines 17–21 for *goretj*. What is also important that the patterns of metaphorical uses are shared between English and Russian *to glow with desire* is naturally rendered as *goret’ zhelaniem*. The study in the second part of the book shows that the situation is similar with other words as well, i.e. English, German and Russian share significant amount of patterns of metaphorical uses. This is another example of the Worfian SAE hypothesis. In the network we present the opposition of the two cases as *[congruent]* vs. *[noncongruent]* realisation. In the case of some lines it is necessary to consider the larger context to determine the type of reading. For instance, the larger context from line 7 for *burn*:

(4.14) *He plunged into print-making and in too short a time had effectively burnt himself out.*

suggests the noncongruent reading.

There are two types of *[congruent]* reading: the production of *[light]* or *[flames]*. Note that this distinction does not imply that there are two separate concepts behind the uses of the two verbs. The opposition can be described as a resource for the speaker who can highlight either of the aspects of one and the conceptual representation and render it in different ways. The choice of either option also does not prescribe the choice of the verb. Many uses of *glow* refer to the production of light, many uses of *burn* refer to flames, but the opposite is also possible, e.g. line 12 of *burn*.

Another distinction concerns the agency and transitivity. The verb can correspond to a directed material process (the one that extends over another participant in the clause) or to a non-directed material process. In the former case there is a possibility of an explicit or implicit agent, as in lines 3 vs. 14 of *burn*, respectively.

The choice of *[light]* entails a choice between two further options:

1. the emphasis on light itself *[luminescence]* or its colour *[colour-specification]* and

2. the difference between the light produced by the glowing body or reflected by it

---

6The terminology used here follows (Halliday, 1985) and (Halliday and Matthiessen, 1999)
4.4. CASE STUDY: THE DESCRIPTION OF “TO BURN”

Similarly a reference to [flames] entails a choice between two other options:

1. the emphasis that an object is being destroyed by fire [destruction] or this is part of its operation [nodestruction] and
2. whether the fire has been set deliberately or not: [deliberate] vs. [nondeliberate], the former case may include the actor.

Options referring to the [noncongruent] reading are much more diverse and the short set of examples does not allow the construction of a coherent set of options. The most developed network concerns the expression of feelings. This includes, in particular,

1. the expression of [pain] as burning, e.g. lines 2, 6 10 for burn. The Russian lines do not list such uses, but they are also possible, e.g. vnutili nap menja vsjo gorelo. This is a universal transfer grounded in human experience, when the stinging sensation caused by the fire is transferred to similar feelings in the body.

2. [emotions], they can be either [positive], [neutral] or [negative], as discussed above.

3. [emptiness] is used for classifying two deviant lines of to burn: 7 and 15, which also refer to a feeling, namely to the feeling of one’s inability to continue normal operations as the effect of “burning”. A larger set of concordance lines and verbs of burning in the two languages might reveal regularities behind the uses.

Other options involve references to [end-of-service] (lines 9 for burn and 4 for gotet’), as well as Russian only features [expectation] (lines 1, 2, 7, 11, 18 for svetit’), [showup] (lines) or (lines). At the same time noncongruent realisations cannot be treated as completely separate cases, because they are typically closely related with other areas of the network, for instance, emotions, see Chapter 7, and can be best studied in relationship to them.

This informs the systemic network shown in Figure 4.6.

Some lexical item are inherently related to features in the network. For instance, gotet’ is always related to [nondeliberate] happening. Even though to burn follows same patterns of uses as the Russian gotet’, it can be used in be used both in the context of [deliberate] and [nondeliberate] fire, cf. lines 13–17 to lines 18–20.
CHAPTER 4. THE LEXICAL MODEL

Figure 4.6: The basic options for *burn*
4.4. CASE STUDY: THE DESCRIPTION OF “TO BURN”

There are some concordance lines that do not allow an interpretation in the network. One of them is line 5 for burn, which refers to the execution by burning a person alive. This is a congruent realisation of the concept, i.e. physical destruction, and it is systematically related to other senses of this verb. One solution would be to add another subtype of physical burning along with [light] and [flames], however, there is no clear relationship of execution to other options available in the system. This also shifts the perspective of natural combination between options. So from the viewpoint of communicative purposes it is more natural to represent this type of uses in another network, which describes types of killing and respective expressions, including the verbs kill, murder, strangle, etc. It is interesting to observe the communicative difference between two conceptually similar events. Line 5 in the English concordance for burn refers to the execution (of Jan Hus), whereas line 10 in the Russian concordance for goret’ mentions another execution (of Giordano Bruno). However, the latter example refers to the event not from the viewpoint of the execution (this could be rendered as sozhen na kostre), but from the viewpoint of the effect this event had on further research: the use of this word actually disapproves Bruno’s nonconformism by comparing it to Galileo’s conformism, who continued to work even though he had to denounce his views. Thus, the word choice (sgorel) in this case realises the communicative intention similar to sgoret’ in line 7, which mentions that the house of Enron’s CFO has been burnt in Houston.

The other concordance lines are fully represented. The network looks quite complicated, especially given its purpose to describe just four words in two languages. However, the growth of the network is similar to a steep learning curve. The initial attempt to arrange few examples in terms of general categories underlying their uses leads to many options. However, once basic choices have been made, they can be used for other purposes. The network is broad enough to cover uses of other verbs in the lexical domain of burning, for instance, to ignite, which is associated with such features as [cause] and can realise either [flames] or [emotion] features. Another verb in the domain shine, which can realise such features as [luminescence] [bright] or also [emotion] [positive] (eyes shining with enthusiasm). The network in Figure 4.6 touches also the domain of expressions available for expressing emotions. This part of the network might be linked with the one that is described in Section 7.4.

What is more, the choices required for distinguishing between separate senses of burn and flame are applicable to other lexical domains, for instance, the aspectual characteristics of the process ([cause], [process] or [effect]). Similarly, the general distinction between a congruent realisation of
a material process referring to an observable event and its reinterpretation in other domains is relevant for many other events, as it is confirmed by the study of the lexical domains of size adjectives and verbs of motion in the second part of the book.

4.4.3 *burn*: the view from below

This view maps uses of *burn* in concordance lines to features in the network. Every use instantiates a set of features, which can be assigned in the incremental fashion using the interface of the Systemic Coder (O’Donnel, 2002). For instance, coding of line 1 for *burn* from Figure 4.4 prompts the choice between the generation of light or flames after choosing the option of
4.4. CASE STUDY: THE DESCRIPTION OF “TO BURN” 103

congruent reading (Figure 4.7). In a real study the process of coding often informs the choices in the network, so the initial network consists of basic choices available a priori and is gradually rearranged in the process of coding to accommodate for examples from concordance lines. In the end every line gets an analysis and their totality can be investigated for the relationship between senses and languages. The results for some of the lines from Figure 4.4 are presented in Table 4.1 (figures correspond to concordance lines, whereas figures in parentheses correspond to numbers of examples in the book).

| congruent light production luminescence process | burn (2.8); glow 4, 9, 17; goret’ 14, 22; svetit’ 3, 6, 9, 10, 12, 14–17, 20. |
| noncongruent feeling emotion positive explicit process | glow 20–25; goret’ 21; svetit’ 13 |
| congruent flames destruction effect | burn 8, 18, 20; goret’ 1, 7, 11 |
| congruent flames destruction process | burn 2.9; goret’ 2, 3, 8, 12, 13, 16 |
| congruent light luminescence reflected process | glow 10, 14, 16; goret’ 15; svetit’ 4, 8 |
| noncongruent feeling pain effect | burn 2, 6, 10 |

Table 4.1: The results for burn and glow

The proposed network is similar to the network drawn in (Hasan, 1987). both accounts belong to the systemic approach and adopt the same formal mechanism. The subtle differences (which are often the major cause of factional fights) are related to two issues. The first issue concerns the corpus orientation of the proposed model. The description takes the most frequent words in the lexical domain related to burning and considers all their most frequent meanings. Hasan’s description focuses on just a few words taken in their single sense (some of them are also rare, like strew). It is also based exclusively on the intuition. The proposed description utilises the intuition and the structure of senses in dictionaries, however, it tries to cover all the frequent types of uses in concordances. The corpus orientation of the case study of burn is not complete, because it is based on the small set of concordance lines shown in Figures 4.4–4.5. A study based on a larger corpus will definitely bring some other issues related to the use of the words in question. The studies reported in Part II are designed to stand a stronger claim: namely that they provide a complete coverage of the lexical domain,
because they are based on a study of a much larger and diverse set of corpus resources, not only concordance lines, but also collocates and other available linguistic descriptions.

Second, the proposed description makes no claim with respect to the delicacy of description, while Hasan’s model assumes that lexis is most delicate grammar, so the lexical choices follow the grammatical ones. Even though the proposed description is confined to verbs, so all lexical choices may be treated as more delicate than the design of the clause in terms of Process and Participants. At the same time the Process can be realised as either material or mental depending on the choice between [congruent] or [noncongruent] realisation. In this respect, the proposed model is similar to that of Gordon Tucker (1998). What makes it different from Tucker’s model is the scope and greater functional orientation: his description covers only the single literal sense of adjectives and does not relate them to the functions of their uses.

As for problems like those in (2.8), when word use does not fit the set of necessary and sufficient conditions for identifying the sense as a concept, they do not arise in the proposed communication-centred description, simply because it does not assume the existence of a concept underlying the word use.

In the rest of the book we are going to study the lexicogrammatical cartography of three domains: size adjectives, verbs of motion and words referring to emotions (with the emphasis on the lexicon of anger and anxiety). The three studies involve larger groups of lexical items than just burn and glow, but they are still incomplete. The complete lexicogrammatical cartography of the size of the coverage of WordNet is a much grander task that can be accomplished a single-man job. The first two studies consider the variety of purposes for using words from the domain, such as little brother, but without studying words from other domains that can be used to express size or motion in a suitable context. On the other hand, the study of the lexicon of anger and anxiety takes the opposite view: it considers a set of lexical items in the domain without studying their possible uses in other domains, such as mad.
Part II

The Comparative Lexicogrammatical Cartography
Chapter 5

The communication-centred approach in action: uses of size adjectives

5.1 The methodology for description

There is a natural balance between what is helpful and what is feasible. A study of just two words in just two languages on the basis of about 20 concordance lines, such as the study reported in Section 4.4, leaves little possibility to make any general claim about the functions of uses, because the lexical resources of any given language have a large number of potential expressions for communicating the same or similar situation. A study of two words in each language leaves many dead links leading to nowhere. For instance, Russian translations of two lines from Figure 4.4

\(5.1\) . . . \textit{Jan Hus preached before being burnt for heresy in 1413}

\(5.2\) \textit{You can burn just as easily on a cloudy day}

use verbs \textit{szhechq} and \textit{obgoretq} respectively. The verbs clearly belong to the same semantic field, but they are not covered in the description. A study of lexical resources in two languages is also limited, because of the number of idiosyncratic expressions. If we want to find regular patterns available in lexicogrammatical resources, we have to extend the range of languages in order to balance linguistic idiosyncrasies.

At the same time, a large-scale study has its own limitations. Development of a new dictionary (even covering the core lexicon) is an expensive
enterprise. It requires the study of many thousands of concordance lines. A careful lexicographical description also requires native or near-native proficiency in languages involved. This restricts the possibility of a typological study covering dozens languages.

Out of these reasons, the studies presented in the following chapters are confined to a relatively closed semantic domain and to the three languages: English, German and Russian with the results of studies for English and German confirmed by native speakers and the evidence from general purpose corpora. A lexical domain is never completely closed because of the problems of the form-meaning interface discussed above. The side effect for translation studies is that it still leaves many dead links, but the description has to cover the majority of cases. The experience with describing size adjectives, verbs of motion and words expressing anger shows that the set of about 20-50 words can be sufficiently closed, representative and yet allowing a relatively concise description.

A lexical domain can be described in the following steps:

1. compilation of a list of most frequent lexical items in the domain (the notion of and the frequency is based on data from available corpora);

2. detecting basic choices for expressing events, objects, and their properties in the domain; the result is represented by the systemic choice network and conditions for making choices in it;

3. checking how the basic choices cover the existing descriptions in available sources, i.e. mono- and multilingual dictionaries and lexicographical descriptions, and detecting most important contexts of their use;

4. extending the network of choices to cover all possible uses of the lexical items under consideration.

In terms of the three viewpoints on the lexicogrammar discussed above, the methodology for description is situated “within” the lexicogrammar, but it should take into account the viewpoints “from above” and “from below”. We start our exploration with concordance lines (the view “from below”), which are then classified into basic types of uses to create the systemic network (the view “within”) and relate them to the functions underlying their use (the view “from above”).

On the one hand, the network resulted from the study is limited to words in the domain, more precisely to their regular uses. On the other hand, it includes parts that might be relevant for other descriptions. For instance,
the description of size adjectives presented below is a part of the lexicogrammatical network which describe qualities of objects. In particular, it does not cover the realisation of terms, which is left out, because the realisation of terms should be modeled outside of the network of quality, as it pertains to the level of the nominal group specification. Even though a big toe carries its name for a reason (it is bigger than other toes), its realisation should be specified in the same part of network that specifies thumb designating the respective finger. At the same time, the network includes portions that are shared with networks describing other words. It is quite probable that portions of the network for size adjectives are relevant to networks for other types of adjectives. For instance, the description below hints the importance of evaluation in the choice of size adjectives. However, many uses of size adjectives do not contribute to evaluation, so the description of lexical means for evaluating qualities of objects will result in a much more developed network.

Sections below follow the steps in the methodology for describing the lexicographic domain of size adjectives.

### 5.2 Compilation of the list of size adjectives

Table 5.1 lists the size adjectives to be considered in English, German and Russian. The study considers all the size adjectives that belong to a frequency list that would cover more than 75% of word uses in a representative corpus. The threshold of 75% is also used in CCED (1995) as the definition for the core lexicon (such entries are marked with 5 or 4 diamonds in CCED). The current estimation is that about 2000 words in a language are within the 75% threshold.

The English set is based on frequency data from the word list compiled by Adam Kilgarriff on the basis of the British National Corpus (Kilgarriff, 1996). Words starting with the frequency lower than about 60 ipm in the BNC are outside of the 75% threshold according to CCED (1995), however, they were included in the list, because they are important for the sake of better coverage of the domain. For instance, both thin and thick are beyond the threshold, however, it is clear that they are important members of the list of size adjectives. Similarly, shallow is much less frequent than deep; their respective ranks in the BNC are 15.13 vs. 84.44 ipm. Most frequently, small degrees of depth are addressed in English by means of negation, e.g.

(5.3) *There is a very lovely lough close by called Sweethope, . . ., not very deep and full of splendid trout.*
<table>
<thead>
<tr>
<th>English</th>
<th>Rank(freq)</th>
<th>German</th>
<th>Rank(freq)</th>
<th>Russian</th>
<th>Rank(freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>great</td>
<td>138 (644)</td>
<td>long</td>
<td>59 (687)</td>
<td>tol’shoj</td>
<td>62 (1618)</td>
</tr>
<tr>
<td>high</td>
<td>178 (527)</td>
<td>wenig</td>
<td>64 (588)</td>
<td>malen’kij</td>
<td>229 (396)</td>
</tr>
<tr>
<td>small</td>
<td>183 (516)</td>
<td>groß</td>
<td>90 (337)</td>
<td>vysokij</td>
<td>294 (303)</td>
</tr>
<tr>
<td>large</td>
<td>198 (472)</td>
<td>hoch</td>
<td>117 (211)</td>
<td>ogromnyj</td>
<td>308 (295)</td>
</tr>
<tr>
<td>long</td>
<td>230 (405)</td>
<td>weit</td>
<td>177 (127)</td>
<td>velikij</td>
<td>376 (252)</td>
</tr>
<tr>
<td>big</td>
<td>262 (333)</td>
<td>klein</td>
<td>252 (74)</td>
<td>dünnij</td>
<td>416 (257)</td>
</tr>
<tr>
<td>little</td>
<td>370 (270)</td>
<td>kurz</td>
<td>451 (38)</td>
<td>korolkij</td>
<td>504 (198)</td>
</tr>
<tr>
<td>low</td>
<td>376 (266)</td>
<td>knapp</td>
<td>566 (30)</td>
<td>shroksij</td>
<td>557 (181)</td>
</tr>
<tr>
<td>major</td>
<td>426 (236)</td>
<td>gering</td>
<td>714 (23)</td>
<td>neskol’shoj</td>
<td>566 (178)</td>
</tr>
<tr>
<td>short</td>
<td>588 (176)</td>
<td>eng</td>
<td>876 (19)</td>
<td>tolstij</td>
<td>626 (163)</td>
</tr>
<tr>
<td>wide</td>
<td>649 (156)</td>
<td>breit</td>
<td>1207 (14)</td>
<td>krupnjij</td>
<td>649 (159)</td>
</tr>
<tr>
<td>fine</td>
<td>697 (146)</td>
<td>niedrig</td>
<td>1212 (14)</td>
<td>tonskij</td>
<td>767 (138)</td>
</tr>
<tr>
<td>deep</td>
<td>1197 (84)</td>
<td>tief</td>
<td>1245 (14)</td>
<td>dolgij</td>
<td>820 (131)</td>
</tr>
<tr>
<td>huge</td>
<td>1250 (79)</td>
<td>dick</td>
<td>- (60)</td>
<td>глубокий</td>
<td>825 (131)</td>
</tr>
<tr>
<td>broad</td>
<td>1488 (64)</td>
<td>fein</td>
<td>- (58)</td>
<td>melkij</td>
<td>861 (124)</td>
</tr>
<tr>
<td>tiny</td>
<td>1701 (56)</td>
<td>rieseng</td>
<td>- (55)</td>
<td>szkij</td>
<td>1036 (90)</td>
</tr>
<tr>
<td>thin</td>
<td>1715 (55)</td>
<td>nieder</td>
<td>- (40)</td>
<td>niskij</td>
<td>1470 (70)</td>
</tr>
<tr>
<td>narrow</td>
<td>1772 (53)</td>
<td>grok</td>
<td>- (30)</td>
<td>gromadnyj</td>
<td>2706 (40)</td>
</tr>
<tr>
<td>tall</td>
<td>1781 (53)</td>
<td>maßig</td>
<td>- (23)</td>
<td>tsynj</td>
<td>3137 (34)</td>
</tr>
<tr>
<td>brief</td>
<td>1861 (50)</td>
<td>schmal</td>
<td>- (23)</td>
<td>men’shij</td>
<td>3739 (28)</td>
</tr>
<tr>
<td>minor</td>
<td>2014 (45)</td>
<td>dünn</td>
<td>- (22)</td>
<td>nevyysokij</td>
<td>3942 (26)</td>
</tr>
<tr>
<td>thick</td>
<td>2692 (44)</td>
<td>dick</td>
<td>- (22)</td>
<td>nektlozhny</td>
<td>4320 (24)</td>
</tr>
<tr>
<td>slight</td>
<td>2711 (38)</td>
<td>dünn</td>
<td>- (22)</td>
<td>nevyysokij</td>
<td>3942 (26)</td>
</tr>
<tr>
<td>shallow</td>
<td>2448 (15)</td>
<td>dünn</td>
<td>- (22)</td>
<td>nevyysokij</td>
<td>3942 (26)</td>
</tr>
</tbody>
</table>

Table 5.1: The most frequent size adjectives in English, German and Russian
5.3 Detecting basic choices

The set of lexical items identified in Table 5.1 suggests the following three basic oppositions for using size adjectives in all the three languages: dimension, reference and emphasis (cf. the network in Figure 5.1). The choices of respective features are done in parallel; then the [directional] feature for dimension offers more delicate choices.

**The dimension of measurement:** [non-directional] vs. [directional]

This option checks the applicability of size adjective to a particular direction of an object. If [non-directional], this corresponds to such adjectives as *big, small* or *slight*. If [directional], the following options can be considered: along object’s length (*long, short*), width (*wide, narrow*), thickness (*thick, thin, this choice is directional, but different from width*) and height (in this case, the view on the measurement is important: it can be directed upwards from below, *high, low*, or downwards from above, *deep, shallow*).
The reference of measurement: [small] vs. [big]
This is an option available for all types of measurements: an object can be either small or big, long or short, thick or thin. The bigger side of the scale is slightly more frequent, i.e. typically high is more frequent than low, groß than klein, bol’shoj than malen’kij, though small is more frequent than large.

The emphasis type: [emphasized] vs. [non-emphasized]
This opposition provides the possibility to choose between large and huge or between little and tiny. Another possibility concerns non-lexicalised choices for [emphasized], which are expressed in the grammatical structure of the quality group: long vs. very long.

The proposed basic classification can be applied to encoding the size adjectives. An example of feature traversal for English adjectives is given in Table 5.2

<table>
<thead>
<tr>
<th>English</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>big</td>
<td>big non-emphasized non-directional</td>
</tr>
<tr>
<td>deep</td>
<td>big non-emphasized directional height down-height</td>
</tr>
<tr>
<td>high</td>
<td>big non-emphasized directional height up-height</td>
</tr>
<tr>
<td>huge</td>
<td>big emphasized non-directional</td>
</tr>
<tr>
<td>large</td>
<td>big non-emphasized non-directional</td>
</tr>
<tr>
<td>little</td>
<td>small non-emphasized non-directional</td>
</tr>
<tr>
<td>long</td>
<td>big non-emphasized directional length</td>
</tr>
<tr>
<td>low</td>
<td>small non-emphasized directional height up-height</td>
</tr>
<tr>
<td>narrow</td>
<td>small non-emphasized directional width</td>
</tr>
<tr>
<td>short</td>
<td>small non-emphasized directional length</td>
</tr>
<tr>
<td>small</td>
<td>small non-emphasized non-directional</td>
</tr>
<tr>
<td>thick</td>
<td>big non-emphasized directional thick</td>
</tr>
<tr>
<td>thin</td>
<td>small non-emphasized directional thick</td>
</tr>
<tr>
<td>tiny</td>
<td>small emphasized non-directional</td>
</tr>
<tr>
<td>wide</td>
<td>non-emphasized directional width big</td>
</tr>
</tbody>
</table>

Table 5.2: Annotations of English adjectives

5.4 Testing the basic choices
Even though the description in the previous section is based on the intuitive understanding of what size adjectives are, it fails to take into account
the complete range of possible uses of size adjectives and does not make a clear picture of the communicative purposes they are used for. The existing studies of size adjectives, for instance (Apresjan, 2000:236ff), (Bierwisch and Lang, 1989), (Tucker, 1998) also concentrate mostly on physical features of respective objects without an attempt to relate them to the whole range of uses of size adjectives. As it is often the case with most frequent words, the size adjectives listed in Table 5.1 have a large number of polysemous senses in dictionaries. For example, great has 12 senses in CCED and 9 in WordNet; high has respectively 22 and 12 senses. Analogously, groß has 69 senses in Wahrig\(^1\) and 18 in Duden; bol’shoj has 15 senses in Ushakov and 8 Ozhegov. In total, the lexicographical database of size adjectives lists 365 senses for the 66 size adjectives. The corpus study also confirms that their non-spatial uses are statistically more frequent than strict spatial ones (see the most significant collocations below). In the following subsections we consider specific senses of size adjectives in the three languages, especially the differences in uses of almost synonymous adjectives.

### 5.4.1 English

Size adjectives show the same set of problems of treating lexical meanings as concepts. WordNet (version 1.6) lists 11 synsets for little, including:

1. limited or below average in number or quantity or magnitude or extent;
2. of little importance or influence or power; of minor status;
3. (informal terms) small and of little importance;
4. contemptibly narrow in outlook, e.g. "a little mind consumed with trivia"; "petty little comments";
5. used of persons or behavior; characterized by or indicative of lack of generosity.

It is clear that synsets 5, 6, 8, 11 have something in common, namely, the negative attitude towards the object designated using little in this sense, but there are no formal means to relate them in the definition. Also, it is not always easy to select the exact concept in the lists of senses. For instance, little is used twice in the following example:

\(^1\)(Wahrig, 2000) sometimes enumerates possible uses, including terminological collocations as separate subsenses (like die große Armee, the army of Napoleon in 1812). Thus, the number of senses in Wahrig is much larger than in other dictionaries and they were not included into the database.
(5.4) – Oh, you wicked little thing! – cried Alice, catching up the kitten, and giving it a little kiss to make it understand that it was in disgrace.

The first use (little thing) can refer simultaneously to synsets 1, 5, 6 and 8, even though the example is not ambiguous. Small also has 11 synsets, small\textsubscript{1} equals to little\textsubscript{1} while small\textsubscript{2} = minor\textsubscript{10}, which is described as “limited in size or scope”, which is almost indistinguishable from small\textsubscript{1}. At the same time, there is no appropriate sense in the WordNet list to classify a little kiss, the second use of little in (5.4), which in principle can refer to “below average in magnitude”, though there is no unit to measure the magnitude of kisses.

The list of synsets and its relationship to words is language-specific. For instance, English provides an array of words referring to the small-scale generic size specification: little, small, slight, and minor. They are typically translated into German as klein (small). However, we cannot assume that klein shares synsets with the four English adjectives, because German has another array of words referring to small amount specification: gering, klein, knapp, maßig, wenig. Similarly, in Russian another array of words is used for the same purpose: malenkiй, малый, мелкий, небольшой. The words in the German and Russian lists have different uses, which do not fit nicely into the synsets available for the four English words. For instance, малый collocates with kolichestво (small amount), predprinimatel’stvo (small business), ploschad’ (small area), sily (little efforts), skorost’ (low speed), vysota (low altitude), moschnost’ (low power), srok (short term), etc, but WordNet has no synsets that include both small and low. This requires several separate Russian synsets, which differ from synsets available for English.

Unlike WordNet, which aimed at listing all the senses possible for a word, other studies of size adjectives focused mostly on physical properties of objects designated by size adjectives. For instance, Tucker (1998:138) proposes a systemic network for describing size adjectives starting with four features: three dimensions (height, length and width) and generic size references. However, the description is not complete. First, it makes no reference to non-spatial senses, including those for specifying little in the example (5.4) and non-dimensional uses, like great, major, etc. The network also misses several important dimensional adjectives that are listed in Table 5.1: broad, deep, thick, thin. Finally, it does not address restrictions on spatial uses of size adjectives. Every physical object has three dimensions, so one can describe it in terms of its length, width and height. However, not all references are equally possible. A house, as a physical object can be referred to as long
5.4. TESTING THE BASIC CHOICES

or high/low, but it is almost never referred to as short (at least, there is no instance of a short house in either the BNC or the Bank of English). At the same time, a lying pole can be described as long, while the same pole, when it is in upright position, is either long or high (here, both long and high refer to the same dimension). On the other hand, even though, the shape of a cigarette closely resembles the one of a pole, it is impossible to refer to a cigarette as high, even it is held vertically. And vice versa it is impossible to say *long tower. These restrictions come from the “normal” functional position of objects. The direction of the management is also important: a rope attached to a high point shares all the spatial properties of a pole and the hanging position is perfectly functional, but such a rope cannot be regarded as high in English. Width is typically defined as the second dimension (the assumption is that the largest dimension is referred to as length), but there are many cases, when wide refers to the largest dimension, as in wide table.

At the same time, in spite of some problems with covering the set of real uses, the network of size adjectives proposed in (Tucker, 1998) is useful as the starting point for the investigation of their uses. It was extended by gathering all potentially relevant frequent size adjectives (listed in Table 5.1) and by studying respective monolingual corpora, as described below.

There are two pairs of English size adjectives that are considered synonymous in many dictionaries: big vs. large and little vs. small. For instance, WordNet contains many synsets (sets of synonymous senses) in which both adjectives are presented, e.g. small, little (limited or below average in number or quantity or magnitude or extent); insignificant, little, peanut, small (of little importance or influence or power; of minor status); little, small (very young), etc. However, the adjectives in the pairs have quite different patterns of usage. The following nouns are the most significant right collocates for them:

big: advantage, apple, band, bang, bank, ben, boss, boy, break, bream, brother, business, car, cat, challenge, chance, city, club, company, crowd, day, deal, difference, disappointment, event, family, fan, firm, fish, four, game, girl, gun, hit, hole, house, increase, issue, job, mac, man, match, mistake, money, mouth, name, part, picture, piece, prize, problem, question, race, room, screen, shock, star, step, success, surprise, thing, toe, trouble, wave, way

---

2The example is from (Dirven and Talyor, 1988) following an analysis from (Bierwisch, 1967).

3For some pairs of words, the T-score of their collocation is shown. For example, small town or little girl are much stronger collocates than little town or little girls, respectively.
large: amount, animal, area, audience, block, body, bowel, bowl, building, chunk, city, collection, company, corporation, crowd, degree, dose, estate, extent, family, firm, fish, garden, group, hole, house, increase, majority, market, measure, number, organisation, organization, pan, part, party, percentage, piece, population, portion, proportion, quantity, range, room, sample, scale, section, share, size, slice, stone, store, sum, tank, town, tract, unit, volume

little: attention, baby, bastard, bird, book, box, boy (27.8), brother, bugger, chance, change, child (10.7), choice, creature, difference, difficulty, dog, doubt, effect, effort, evidence, finger, girl (33.6), group, help, hope, house, idea, impact, information, interest, kid, knowledge, lad, man, money, opportunity, piece, point, progress, prospect, research, resemblance, rock, room, scope, sense, shop, sign, sister, smile, success, support, sympathy, thing, thought, time, town (10.6), use, value, village (10.0), wonder (10.9)

small: amount, animal, area, bird, boat, bowel, boy (17.7), business, businessman, car, change, child (19.9), community, company, country, farm, farmer, firm, fish, fortune, fraction, garden, group, hole, hotel, hour, house, increase, intestine, investor, island, mammal, minority, number, part, party, percentage, piece, print, proportion, quantity, room, sample, scale, shop, size, square, step, sum, table, team, town (23.6), unit, village (13.8), voice, window, wonder (10.5)

The list of collocations shows that the size adjectives are quite frequently used not for measuring the size of the object itself (it is also hard to compute the physical or metaphorical size of many “objects”, such as attention or surprise), see also the random set of concordance lines for large in Figure 5.2. In this respect, the collocation pattern separates large and small from big and little. Large and small typically refer to the size (large area, fish, room), while big and little are typically used for designating the fact that an object belongs to the class of big things or events (big game, players, surprise). Even, cases, when large, small vs. big, little collocate with identical lexical items (city, company, eyes, fish, house, money, part, screen), they often refer to different characterisations of the same object: a big fish belongs to the class of big fishes (also, figuratively), while a large fish refers to the size of a particular fish.

The system of height measurements is also not trivial. High and tall have different collocational properties. Tall in opposition to short is applied to persons, while it lacks a true antonym, when applied to objects, partial

---

4The presence of medical terms can be explained by the “whelks” problem: the scientific part of the BNC is based on a medical journal (Kilgarriff, 1997b).
accessible and visible to the public at large and concluded that there had been no large scale ecological catastrophe. None of the large European banks have been heavily affected by the financial crisis.

and small countries alike. A systematic approach to protect individuals society at large and the environment against the harmful effects of large amounts of wood in the forest and from international support will depend on a large extent on the way non-Albanian minorities will be treated.

Furthermore, there is a large degree of state protectionism in the French and Italian markets. There is a large degree of state protectionism in the French and Italian markets. There is a large degree of state protectionism in the French and Italian markets. There is a large degree of state protectionism in the French and Italian markets. There is a large degree of state protectionism in the French and Italian markets.

Moreover, the call to make a cultural and political move towards a large extent that this will be reflected in the way non-Albanian minorities will be treated.

The international support will depend on a large extent on the way non-Albanian minorities will be treated.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

In conclusion, the question which concerns an exceptionally large number of working doctors throughout Europe remains unanswered.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

The fact that in Austria the will of the large majority of the people is being respected by the national authorities is crucial.

Figure 5.2: The concordance for large
antonym is low, which is also an antonym to high. Tall can also be treated in opposition to small (tall tree/small tree). Dirven and Taylor conclude that the most general meaning of tall is the designation of vertical objects with one prominent dimension and that it exhibits a polysemy between two concepts: (1) extreme prominence of vertical dimension, bulky, massive, e.g. tall mountain, and (2) extreme prominence of vertical dimension, thin, slender, e.g. tall pillar. However, even though they claim that the intuition of native speakers supports the application of tall to bulky and massive objects (like mountain), the most frequent nominal collocations of tall are nouns referring to persons, trees and windows, while mountains are described almost always as high (we were unable to find uses, in which tall is applicable to bulky objects better than high).

The speaker makes a reference to a property of an object not only for ideational reasons, but also for providing a rhetorical impact onto the hearer. Both the small and big sizes can be associated with positive or negative attitudes depending on the nature of the object being described. The reference to the size of an object/person may be mentioned to justify the need in taking care of it (small, sympathy) or being cautious of it (big, antipathy). On the other hand, importance is typically described in the opposite way.

The speaker’s evaluation of an object influences the pattern of uses of roughly synonymous lexical items. WordNet 1.6 defines a synset for denoting young children:

\[ \text{little}=\text{small}=\text{young} \] (not fully grown, )

However, little is typically used, when the author announces the positive attitude to a child being described. On the contrary, when small is followed by such words as boy, it is typically used in less favorable contexts. Compare the two examples (5.5) and (5.6), in which little and small cannot be swapped over:

(5.5) My little boy is extremely happy there and is coming on by leaps and bounds.

(5.6) A small boy and a begrimed, bowlegged toddler lurked behind them.

This is also confirmed by the study of collocations. Little collocates with girl much more strongly than with boy, while boy collocates stronger with small, cf. Table 5.3, in which the first two columns present the number of the joint occurrences of two words, and the last column presents the total number of occurrences of the respective nouns in the BNC. The difference in the collocation frequency does not necessarily imply the difference in semantics of words boy and girl. However, the difference is based on the
fact that girls are more often referred to in the paternalistic tone, i.e. the word *girl* has the greater probability to be used in conjunction with the interpersonal meaning of sympathy. Thus, it is more often realised with *little* than with *small*. Also, the plural noun forms for *girls* and *boys* correspond to about 61% of uses of the singular form, but the plural compound *little girls* corresponds to just 24% (21% for *boys*) of respective singular uses. There is no such a big drop for *small* (about 50% for both *boys* and *girls*).

This fact can be also explained by the fact that the greater interpersonal value is expressed in the exchange about a particular child, not about children collectively. The difference between patterns of uses of *little* to predominantly formal contexts in which *small* is used (*pretty little girl* vs. *comparatively small quantity*).

However, the distinction in the interpersonal attitude is no absolute. It is not usual in the English culture to refer to small children in a negative context, so typical examples of uses of *a small boy* are like *When I was a small boy; He had two small boys*, etc. At the same time, *little* can occur in apparently negative contexts, like *a silly little boy*, but even this use still keeps the positive general attitude towards the child, by being affectionately patronising.

Another difference between *little* and *small* concerns the opposition of spatial vs. temporal measurements: when *small* is used, it refers more to the size of a child, while *little* to the age. Thus, the joint senses of *little* and *small* in WordNet should be expressed in terms of different network traversals. Also, *little*, when it refers to age, may be used in another function, namely, it can denote the relative age of a person as being younger, i.e. a person may say: *my little brother*, even though he is in his sixties.

<table>
<thead>
<tr>
<th></th>
<th>small</th>
<th>little</th>
<th>BNC count</th>
</tr>
</thead>
<tbody>
<tr>
<td>boy</td>
<td>228</td>
<td>686</td>
<td>13290</td>
</tr>
<tr>
<td>boys</td>
<td>113</td>
<td>143</td>
<td>8054</td>
</tr>
<tr>
<td>girl</td>
<td>50</td>
<td>996</td>
<td>15762</td>
</tr>
<tr>
<td>girls</td>
<td>24</td>
<td>239</td>
<td>9621</td>
</tr>
</tbody>
</table>

Table 5.3: Frequency values for small and little collocating with children.
5.4.2 Russian

The study of size adjectives in Russian corpora confirms that the general rhetorical principle holds in Russian as well as in English: the size adjectives are not only a device for specifying the size of an object, they express speaker’s attitude and are aimed at influencing the hearer. Let’s consider four near-synonymous adjectives in Russian that refer to the small size or amount of something: malen’kij, malyj, melkij and nebol’shoj.

At the first glance, the difference between uses of malen’kij and nebol’shoj is related to the difference in size: nebol’shoj implies a negation of bol’shoj (big), while malen’kij is small par excellence, so nebol’shoj is greater than malen’kij. Indeed, there are cases, in which the use of malen’kij is not possible, because of the relatively large size of respective objects, for instance, on sidel’ na nebol’shoj/?? malen’koy skale, he sat on a little rock.\(^5\) However, the most examples show no specific restriction on the uses of malen’kij vs. nebol’shoj with respect to the size. In the cases like (5.7) and (5.8), they can be exchanged:

(5.7) Ja jasno razlichal tol’ko nebol’shoj uchastok kamennoy poverynosti.
I only could make out clearly a small patch of the stony surface in front.

(5.8) On uvidel malen’kij chernyj pistolet, lezhavshij u menja na ladoni.
He saw the little dark weapon lying in my palm.

However, there are certain restrictions. One of them, like the difference between uses of *small* and *little* in English, is related to the attitude of the speaker. It is much less likely that the speaker will use nebol’shoj for referring to an object that evokes his/her specific emotions. On the contrary, the attitude of the speaker expressed with malen’kij is typically positive, unless other qualities or the noun itself imply the negative attitude, e.g. *malen’kaya zaplevannaja kamorka* (small dirty closet). The use of nebol’shoj instead of malen’kij in the both examples should contrast with the emotional attitude towards the object. On the contrary, nebol’shoj is more felicitous in neutral contexts, e.g. the expression nebol’shoj dokument (small document) is used throughout the Word manual (not malen’kij dokument).

Both adjectives can refer to an interval of time nebol’shoj/malen’kaja ostanovka (short stop), though the difference in the interpersonal attitude towards the interval is still important: malen’kaja ostanovka typically implies

---

\(^5\)As discussed by Raxilina (2000).
an interpersonal message “don’t bother, enjoy your life,” while *nebol’sha* *ja ostanovka* is neutral.

There are also some grammatical constraints on the use of *nebol’shoj*: it is unlikely in constructions with modifiers, like *slihkom* (too), *ochen’* (very), with respect to the size:

(5.9) a mozhet byt’, kljuchik byl slihkom *malen’kij*, to’ko on nikak ne xotel otkryvat’

or the key was too small, but at any rate it would not open any of them.

Even the Word manual uses *malen’kij* in such cases in spite of the fact that the more neutral *nebol’shoj* is preferred in general:

(5.10) *libo graficheskij objekt ochen’ malen’kij*

or the drawing object is too small.

This constraint is related to the fact that *nebol’shoj* is a negation of *bol’shoj* (large). Thus, this states the negation of the fact that a property expressed by the adjective is large, i.e. something is not big, but this does not imply that it is necessarily small. So the smallness of an object described as *nebol’shoj* cannot be emphasised by such words as *ochen’* (very) or *slihkom* (too). This is applicable also to other derived words, e.g. *neshiroki* (narrow), *nevysoki* (low), which are also quite frequent, though beyond the 75% threshold.

There is also another interpersonal difference, which distinguishes their uses from respective [small] adjectives without the negation prefix. When *uzkij* (narrow) or *nizkij* (low) are used, they often include a negative characterisation of a value, e.g. *uzkoe mesto* (bottleneck, lit. narrow place), *nizkoe kachestvo* (low quality). On the contrary, *neshiroki* or *nevysoki* are used, when the goal of the speaker (often in advertisements) is to de-emphasize the negative properties of an object being described, e.g. in comparison to its price: *neshiroki* *kanal sujazi* (narrow communication channel), *skanirovanie s nevoskim razresheniem* (low resolution scanning), *nevysokie tseny* (low or competitive prices).

The difference between the uses of *malen’kij, malyj* and *nebol’shoj* will be clear, if we have a look at the lists of their most significant collocations:

*malen’kij*: bukva, chelovechek, derevushka, devochka, ditja, domik, Edward, figurka, glazok, glotok, gorodok, knizhechka, komnatka, kusochek, lebed’, magazinchik, mal’chik, ostrovok, radost’, rebenok, rost, sobachka, starichok, starushka, stolik, xitrost’
nebol’shoj: chast’, chislo, den’gi, derevnja, domik, doza, firma, gorod, gorodok, gruppa, kolichestvo, komnata, komnatka, kompanija, kuchka, kusochek, otrjad, otstuplenie, partiya, pausa, peredvishka, peredryv, ploschadka, poljana, porcija, predprijatie, prostranstvo, rasstojanie, razmer, rost, summa, tolopa, uchastok, vysota, zal, zapas

malyj: biznes, chast’, ditja, dolja, doza, forma, god (bez malogo let), gorod, kolichestvo, krov’, narod, nuzhda, oborot, predprijatie, predprinimatel’, predprinimatel’stvo, rodina, scena, serija, sila (malymi silami), skorost’, srok, stepen’, taz, teatr, tolka, vysota, xod, zal

First, words with diminutive suffixes (-chek, -chka, etc) referring to something ‘pet’ are abundant for malyj, further confirming the hypothesis that it realises the positive interpersonal attitude. Secondly, malyj and nebol’shoj together correspond to Finally, the word malyj (small) with respect to spatial properties is archaic. In modern Russian, it is typically used in more or less idiomatic constructions:

1. various terminological collocations referring to types of things: predprinimatel’stvo, bizness (both business), mehanizatsija, malye formy; skorost’ (velocity), vysota (altitude), moschnost’ (power), ploschadj (area), while a value in the latter four cases should be indicated by means of nebol’shoj;

2. proper names: toponyms (Malaja Dmitrovka), historical objects (Malij Dvorets),

3. expressions that cover the set of possibilities, e.g. bol’shix i malyx gorodov (big and small towns), dl’ja spiska malogo ili srednego razmera (for a small- or medium-size list). When no coverage is implied, then nebol’shoj fits better (Vokrug Moskvy mnogo nebol’shix/ malyx gorodov, There are many small towns around Moscow),

4. short temporal intervals: malye sroki, e.g. dostavki, stroitel’stva (short term, e.g. delivery, development).

5. the short form is used for designating the insufficient size of objects: pomeschenie malo (the room is cramped), sapogi maly, (shoes are tight)

6. bez malogo (all but a few, almost),

7. malaja chast’ (small part), typically with a restriction: lish’/tol’ko malaja chast’ (just a small part),
As in English, the small amount is often designated using a metaphor for measuring along a dimension: nizkij (low), melkij (small), tonkij (thin), korotkij (short). However, the latter two adjectives occur mostly in terminologically bound collocations. Tonkij (thin) in its uses referring to the imaginary measurement is typically used in the positive sense: um (intellect), rabota (work), sluz (perfect ear), obraschenie (manners), while tolstyj (thick) is not used in such collocations at all.

There is a difference between English and Russian in the relative frequency of size adjectives with the [emphasized] feature. Words referring to the [emphasized small] features are less frequent than in English, but various words referring to the [emphasized big] features are more frequent in Russian: velikij (great), gromadnyj, ogromnyj (both correspond to great and huge).

They also have a more complex system of uses than simple [emphasized big], as proposed in Figure 5.1. First, the short form of velikij is an antonym to the short form of malyj, when the size of clothes does not fit (no reference to [emphasized] in such uses). The uses of fitness of clothes are expressed in specific ways in all the three languages and warrant for a specific feature in the network. Second, velikij most often occur with names of persons to designate their superiority in some respect velikij hudozhnik (great artist), velikij greshnik (great sinner), uroki velikogo kalifornijtsa (the great Californian’s lessons). Another possibility in using it is to refer to the superiority of the situation:

(5.11) v atrmosfere velikoj melanzolii, velikogo otvraschenija.  
_in an atmosphere of great melancholy and disgust._

(5.12) nachali spravljat’ velikij natsional’nyj prazdnik.  
_there was some great national celebration._

(5.13) Ja beskonechno blagodaren i emu, i velikoj organizatsii, kotoruju on predstavlil.  
_I am infinitely obliged to him and the great Institution he represented._

Finally, like malyj, it is extensively used in toponyms, e.g. Velikij Novgorod, and terms, e.g. velikij knjaz’ (the Grand Prince), velikaja kitaeskaja stena (the Great Wall of China). The adjective velikij in all the cases above cannot be replaced by either gromadnyj or ogromnyj.

Like big and large in English, the three adjectives differ in the degree for specification of the size vs. the class of an object. Velikij refers to the fact that an object is a member of a class of things with a high degree of
respective properties. *Gromadnyj*, on the other hand, most typically refers to the very large physical size of an object, while *ogromnyj* is somewhere between them, closer to *gromadnyj*. In the following case, the substitution of *gromadnyj* instead of *ogromnyj* is not felicitous:

(5.14) *S ogromnym smakom ona nachertila marshrut.*  
*She had traced our route with immense zest.*

On the other hand, there are cases of synonymous and quasi-synonymous uses, for example:

(5.15) *velikomu vmerasovomu klanu ryzhevolosyx*  
*the great clan of intra-racial redheads.*

*Krupnyj* is yet another Russian size adjective that expresses the meaning of [non-directional big]. Its typical antonym is *melkij*. Both adjectives can refer to:

1. elements of a quantifiable object: *krupnyj/melkij pesok* (sand), *kupjury* (banknotes), *kapli* (drops);
2. terminological collocations: *krupnyj/melkij skot* (cattle), *remont* (repair), *torgovets* (merchant), *vkladchik* (investor), *proizvoditel’* (producer), etc;
3. importance or influence: *krupnyj/melkij pomeschik* (landowner), *neprijatnost’* (trouble), but only *krupnyj talant* (talent), *uspeh* (success), *melkie podrobnosti* (small details);

As for differences between uses of *bol’shoy* and *krupnyj*, *bol’shoy* is used in a wider range of free constructions than *krupnyj* and most typically refers to the physical size of an arbitrary object. When *krupnyj* refers to the size, it is mostly restricted to certain types of objects, like the size of human persons (especially children), or body parts (zuzy, glaza, kosti), or certain animate objects (*ptitsy, sobaka*). The major groups of terminological collocations with *krupnyj* concern the font size, amount of money (*sdelki, den’gi, bogach, torgovets, vkladchik, aktsioner, burzhuzaziya, proizvoditel’, predprijatija*), or elements of a quantifiable object (*kapli*). Other collocations with *krupnyj* are listed above. One possible generalisation of the uses is that *krupnyj* applies to objects that can grow to reach their large size.
5.4. TESTING THE BASIC CHOICES

5.4.3 German

The same set of considerations applies to German size adjectives: they can be used for a wide variety of purposes, which are quite similar to the set of purposes discussed for English and Russian. The study is slightly limited, because there is no reference corpus for German to study general collocates, but some information can be gathered from dictionaries. For example, (Wahrig, 2000) lists nine basic classes of senses for groß:

1. extended in space, e.g. Haus (house), Hände (hands);
2. extended in time, e.g. Ferien (holidays), Pause (pause), Wanderung (walk);
3. large in number of elements or amount, e.g. Familie (family), Teil (part);
4. strong, high level, e.g. Hunger, Lärm (noise), Hitze (heat), Lust (desire);
5. grown-up, e.g. das Kind ist schon groß (the child has grown up now);
6. significant, Werk (work), Denker (thinker), as well as, events and persons: Armee Napoleons, Friedrich, der Große
7. noble (edel), e.g. Geist (spirit), Mensch (man), Herz (heart);

Thus, like in English and Russian, the size specification naturally extends into the domains of amount, time, importance or influence.

It seems that many almost-synonymous possibilities of English and Russian, like big vs. large, high vs. tall, malen’kij vs. nebol’shoj vs. malyj vs. melkij, do not appear among most frequent German size adjectives. The adjectival groups that require additional treatment are eng vs. schmal (narrow), breit vs. weit (broad, wide) and klein, wenig, gering (small).

Klein most typically refers to the physical size of objects (Dorf, Haus, Raum), their class (Firma, Imbiss, Trost), amount (Auswahl, Gruppe, Teil) or the age of a child (Bruder, Kind). Gering is mostly used for non-spatiotemporal specifications of amount or the number of objects, its most significant nominal collocations are: Menge (quantity), Wahlbeteiligung (participation in elections), Anzahl (number), Gewicht (weight), etc; though there are some uses, like Resonanz (response), Chance, which do not fall exactly into the class of numerical amount. Also, like nebol’shoj or nevysokij in Russian, gering specifies a middle point in the scale between large and
small: the amount is lower than expected or desired, but is not small enough of speaking about it as small. This also has a rhetorical impact: the reader should not assume that the value being described is extremely low. This sense is widely used in the political discourse. Wenig is applicable to temporal uses, e.g. Minute, Tag (day), Woche (week), and numerically measurable quantities, e.g. Geld (money), Meter, as well as the number of objects or events, Ausnahme (exception), Unfall (accident).

In the eng vs. schmal pair, eng is basically applied to object’s width either physical or metaphorical, e.g. Straße (street), Beziehung (relationship), Freundschaft (friendship), Kontakt. In contrast, schmal most often implies the restriction not only on the linear size of an object, but also on its thickness or volume, e.g. schmales Buch (thin book), schmale Lippen (thin lips), schmaler Geldbeutel (thin wallet, i.e. low budget), which cannot cooccur with eng; though schmal can be used as schmale Straße (a narrow street), as a synonym to eng. Also, schmal is significantly less frequent than eng (cf. Table 5.1).

According to (Wahrig, 2000), breit is an antonym to schmal, while the antonym of two-dimensional senses of weit is eng (weit can also mean ‘far’). However, the collocation samples of COSMAS show that breit and eng have similar collocates: Straße, Gasse (both street), Schneise (lane), Raum (space), Auswahl, Angebot (choice), Kreis (a social circle). Finally, the most frequent uses of breit collocate with abstract nouns, mostly denoting the wide range of choices or social basis (of support): Palette, Spektrum, Publikum (audience), Basis, Stützung (support), while the material collocates have no relationship to the volume extension implied by schmal and are less frequent than the abstract nouns, e.g. Streifen (band), Schultern (shoulders), Straße, Gasse.

At the same time, weit collocates with Feld (field, in such examples as Kooperation auf dem weiten Feld der Energiepolitik, cooperation in the broad field of energy policy), Sinn (sense), Umkreis (area), Hose (trousers), Welt (world). Weit, unlike breit, is used in predicative and adverbial uses: die Tür weit öffen (to open a door wide). The most uses of weit are based on a metaphor of free space, which provides the possibility for movement.

5.5 Extending the network of choices

The considerations from Sections 5.4 inform significant restructuring of the network presented in Figure 5.1.

First, we should represent the distinction between spatial and non-spatial
uses of size adjectives, since size adjectives can be used a way beyond pure spatial descriptions, see the measure-type classification in Figure 5.3. The metaphor of representing values as size extends across many domains and languages. This seems to be natural, because the physical size of an object is one of the most evident properties for visual perception, so size properties provide the basis for communicating properties in other domains, in which a parameter can vary in certain respects. A reference to the size in this context implies a mapping from a qualitative state into a value, which represents the measure for the state. Thus, words that designate the mapping are naturally applicable both to sizes and quantities. This is true for the three languages, in which all the most frequent size adjectives considered in the research can be applicable to quantities. However, the opposite is not true, because many adjectives are applicable only to amount, e.g. multiple, insignificant. The chooser for this choice uses the following inquiries:

1. Are we going to talk about the physical size of something? If yes, choose [spatiotemporal].
2. Are we going to talk about temporal properties? If yes, choose [spatiotemporal]. This choice also restricts the set of options in terms directionality, see below.

3. Are we going to talk about the gradable amount of something. If yes, choose [quantity].

Also, as the study of the uses of big vs. large or malyj vs. malen'kij shows, there is a possibility to use size adjectives for specifying the class of objects rather than the physical size or amount. This option is the default choice, when no specific measurement type can be detected or when the choice is idiomatic, i.e. directly follows from properties of a modified noun.

When size adjectives are used for the description of quantity, the natural opposition can be drawn between [count-size] and [mass-size] measurements. The inquiry answers the question: is the object semiotically presented as a set of distinct elements? If yes, the feature [count-size] leads to several possible choices with respect to the nature of the set (things, persons or institutions). The latter choice is not relevant for distinguish between size adjectives in the three languages (no frequent uses were detected, say, specifically for counting things), but this choice is potentially relevant for other languages.

![Diagram of size adjectives classification]

**Figure 5.4: The classification of imaginary quantities**

In the case of [mass-size] measurements, size adjectives refer to quantitatively expressed properties. First of all, there is an opposition of uses pertaining to qualities of persons and qualities of things. For non-personal qualities, there are some specific cases, like intensity, money, voice (small voice, low voice, deep voice), even though reasons for designating specific
voice types by size adjectives should be specified in other regions of the network). The coverage of qualities by size adjectives is not complete; the classification of uses should refer to other domains pertaining to expression of various measurable properties, e.g. force, since the conceptualisation of wind may refer to it in English as high (the size domain) or strong (the force domain), while in German and Russian only the force domain is possible for wind: starker Wind, sil’nyj veter.

Within the spatiotemporal measurements, [spatial-lex] and [temporal-lex] features are distinguished. In language (at least, in the languages under consideration) temporal qualities are often expressed by the same lexical means as spatial ones, e.g. This spans a large period of time, they also share such spatial properties as source and destination, e.g. from April to June. However, as with quantity adjectives, there are specific adjectives to address temporal properties, e.g. brief (stop, outline, suit), old (mother, house, trousers, joke, friend, job, three months old). Temporal qualities can refer either to one’s age or to a time interval. In the former case the distinction is between the age specification in relative terms, e.g. his little brother, i.e. younger than the referent, and in absolute terms, e.g. young. Finally, there can be a difference in referring to the age of a child or an adult: groß in German and bol’shoj in Russian used in the sense “an older child” are applicable to a child, but not to an adult. The duration in the three languages is described in one-dimensional terms, e.g. short, kurz, korotkij, or without a reference to dimensions at all, by means of adjectives applicable only to the duration, e.g. brief, kratkij.

With respect to spatial senses, there are two groups of uses of size adjectives: for referring to the dimension proper (this is the most frequent choice) and to the degree of fitness, when an object (for instance, the human body) fits into the space in another object (for instance, clothes). This often leads to specific lexicogrammatical restrictions, e.g. weit vs. knapp in German, velik vs. mal in Russian (only the short forms are used in this sense); these pairs correspond to loose vs. tight in English. In Russian, this option is also used for coding the spatial uses of the adjectives tesnyj (cramped), prostornyj (spacious). Cramped and spacious in English are beyond the frequency threshold, for instance, narrow or small are often used in situations, in which tesnyj is used in Russian, for instance

(5.16) ... in the mirror reflecting our small Eden.
(5.17) ... v zerkale, otrazhavshem nash tesnyj Edem.
(5.18) ... but the first [of two suits] was too tight and the second too ample.
Finally, the size of specific types of objects can be designated in specific ways. This is represented by the system DIMENSION-OBJECT-TYPE in Figure 5.3. For instance, in English there is a difference between designating the height of persons or animals (tall or short) vs. other objects (high or low).

The second opposition considered in Figure 5.3 concerns the size scale. First of all, the choice of the point on the scale, i.e. [big] vs. [small], depends not only on the objective size of an object, but primarily on speaker’s goals and expectations about the size, e.g. Alice in “Alice’s Adventures in Wonderland” considers herself and other objects as large and small depending on the situation:

(I do hope it’ll make me grow large again.

(approaching the house of the Rabbit): she came upon a neat little house

(approaching the house of the March Hare): It was so large a house, that she did not like to go nearer till she had nibbled some more of the lefthand bit of mushroom.

In German and Russian, there are two possible ways for referring to the smaller side of the scale, thus, the REFERENCE-TYPE system in Figure 5.1 should be more delicate. Some adjectives mean really small, e.g. klein, malen’kij, uzkij. Other adjectives are used for the description of objects that are smaller than their large counterparts, and are below the average or expected value for them; though they do not reach the smaller side of the scale, e.g. gering, mäßig, nebol’shoj, neshirokij. When the option is applicable, the choice is made with respect to several criteria:

1. is it important to distance the description from the smaller side of the scale? If yes, choose [below-average].

2. the size of the object in general: can the object be considered as small at all? This concerns such objects as mountains, boulevards, airports. If no, choose [below-average].

3. the size of an object in the current situation: is it possible to consider the object or amount in the particular context as very small? If yes, choose [little].
Since the choice between [below-average] and [little] in German is available only for [non-directional] adjectives, the first option in the list of inquiries above is multilingual (for the two languages), while others are specific for Russian.

Note the asymmetry between the larger side of the scale and the set of options for the smaller side. No adjectives “large, but not that much” are used in English and German, while in Russian negations of the small-size adjectives exist, e.g. nemalen'kiy, neuzkiy, though they are very rare and mostly used in pragmatically marked contexts (Raxilina, 2000:135ff). The only exception is nemalyj, which is quite frequent (though below the 75% threshold) and can be used for the designation of quantity, e.g. dohody (profits), kolichestvo (amount), chast' (part), or duration: srok (period).

The third system presented in Figure 5.4 concerns the interpersonal attitude towards the object, which size is to be described. The attitude can be positive, negative or neutral. In some cases, there are no hard-wired associations between size adjectives and specific interpersonal attitudes, i.e., language does not provide the speaker with lexical choice and the hearer understands speaker’s attitude on the basis of the context or the modified noun, e.g. big coward vs. big heart. However, as discussed above, the difference can be important for the choice of specific adjectives, when some types of uses are associated with particular interpersonal attitudes. One possible interpretation of size adjectives conforms the metaphor “Good things are big”, cf. (Lakoff and Johnson, 1980:187). For example, in German, the basic senses of groß in (Wahrig, 2000) are ‘bedeutend’ (important, great) and ‘edel’ (noble), which are strongly associated with the positive attitude. Also, some senses of big with positive interpersonal attitude are associated with similar patterns of use, e.g. this is very good of you follows the pattern [very] X of you, as courageous, generous (Hunston and Francis, 2000). The inquiry interpersonal-type-q checks which option is possible. The default choice is [neutral-interpersonal].

The simple classification of dimensions shown in Figure 5.1 (length, width, etc) represents the layman view of the spatial structure, but this view is not directly related to its expressions available in language. For example, a book as any three-dimensional object has length, width and height, however, it is an oversimplification to claim that size adjectives follow them: references to its dimensions as a long/wide/high book are hardly possible; most typically, a long book typically refers to the number of pages in it (actually, to the time required for its reading, as time is treated as one-dimensional).

Even with respect to strictly spatial uses, the classification of dimensions
is not straightforward measurement along dimensions, since the choice also foregrounds specific properties of the object that are significant in the context: a wide table is described along its greater dimension to the right and left, when a person sits at it, not across its greater dimension, cf. (Raxilina, 2000). Finally, the semiotic dimensions are not explicitly bound with the measurement along a specified dimension, because the choice of the dimension is already the result of semanticisation, i.e. semiotic encoding of properties of objects according to the resources offered by the system of language. For example, the physical features of a human being are basically identical for English, German and Russian people, however, in English and Russian, the height is expressed as measured along a vertical dimension (tall, wysoki), while in German it is typically non-dimensional (groß) or one-dimensional (lang).

The updated classification of dimensions presented in Figure 5.5 is flattened (in comparison to the classification in Figure 5.1), i.e. the options referring to [non-directional], and [directional] are joined in the same system, because the reasons for choosing a dimensional or non-dimensional expression for an object are closely intertwined, see the discussion below. Thus, the lexicogrammatical classification, as it is reflected in the form, is relatively simple, but the system of choosers is quite complex\(^6\).

\(^6\)The discussion below follows the suggestions from (Raxilina, 2000:125ff).
The first group of choices checks whether an object has one predominant dimension, which is to be expressed (like ropes, poles, fences along which we go, etc.). If such an object extends vertically, this is relevant for its normal function\(^7\) and it is not a part of another object\(^8\), then the [vertical] measurement is chosen (the direction of measurement is specified in the more delicate system VERTICAL-TYPE). Another possibility for choosing [vertical] is given by the positive response to the question, whether it can be considered as a container (this ultimately leads to choosing [depth-size] in VERTICAL-TYPE). Else, if the object is not shaped as a pen, then it can be long. Otherwise, i.e., if it is like a pen, it can be either long [linear-size] or large [non-directional], depending on the response to the inquiry generic-size-q: whether the size of the object is considered in general or from both dimensions (note that it is very unlikely to talk about long non-pen-like objects as large: ?large dress, hair).

The same inquiry generic-size-q concerns the uses of size adjectives for an object that extends in the two sides in the direction opposite to the face of a person that handles it (this condition is necessary to cover such cases as wide table). We can talk about such objects either as wide [two-dim-size] or large [non-directional].

The last question concerns measurement across the imaginary section of objects with respect to their most prominent dimension, e.g. thick book, rope, slab. If the intention is to describe the cross section, then [volume-size] is chosen. The default choice in this system is [non-directional]. It concerns objects which cannot be described by sufficiently specific shape for choosing a directional adjective. A partial view of the chooser structure is given in Figure 5.6.

The more delicate system of vertical measurements presented in Figure 5.5 allows two directions for [height-size] (from the bottom to the top) and [depth-size] (from the top to the bottom). Note a conceptual difference between the two directions: high is applicable to objects that are vertical and are above a surface, but uses of deep are not constrained only to the measurement down below from the surface, so that shelves, space or borders can also be deep. In short, the distance in the case of deep is measured towards a remote surface, which is not necessarily vertically downwards. The difference is shared across the languages under consideration: deep shelf, tief Regal, глубокая полка. Probably, it is related to human practices of

\(^7\)Cf. the example of *a high cigarette in (Dirven and Tylor, 1988)

\(^8\)As mentioned by Apresjan (2000:236), a pole used as a lightning conductor is long, while the same pole that is used as a standalone flagstaff is high.
Figure 5.6: A partial view of the directional-type chooser in KPML.
5.5. EXTENDING THE NETWORK OF CHOICES

Figure 5.7: A nominal group with a lexicalised size adjective

measurement: the height is assessed by immediate visual observation, while
the assessment of the depth (of, say, a lake) is more obscure; it is often
derived from second-order clues or measurements. Thus, the remote surface,
which is implied by deep, is partly hidden, so as an extension of this mean-
ing, expressions with deep refer to an object located inside of something
or not easily reachable, often in the metaphorical sense: deep forest, crisis,
sympathy.

There is a difference in the selection of dimensions between spatiotem-
poral uses of size adjectives and other types of uses, i.e. [class-property] and
[quantity]. In the first case, the choice is based on the output of respective
inquiries within the [directionality-type]. In the second case, it is based on
conventions grounded in the system of language. For example, a hole can
be deep, because it corresponds to vertical measurement downwards, but
feelings can be deep, because of typical ways for talking about them. In
such cases, the choice of an adjective is determined by the name of an ob-
ject. The size of such parameters as the amount of morale (which can be
high), the selection in the range (which can be wide) or the quality of a sleep
(which can be deep) invokes the metaphor of measurement along respective
dimensions. Metaphors that relate abstract objects to particular dimensions
CHAPTER 5. SIZE ADJECTIVES

are relatively consistent and applicable to many other uses. For example, emotions are considered along a vertical extension. Thus, they can be not only deep, like other vertical objects, they can rise and fall. Also, one can fall into feelings of certain types, like one falls into holes.

5.6 Representing translation equivalence

The contrastive study represented as a systemic network can be also helpful for the tasks of machine translation and multilingual generation. In the case of understanding, the process starts “from below”: each lexical item instantiates a set of features in the network that can be potentially related to meaning intentions responsible for selection of the features. The amount of possible features and meaning intentions is vast, because of high polysemy of size adjectives. For instance, the [little] and [non-directional] features are always selected for little in the REFERENCE-TYPE and DIRECTIONALITY-TYPE systems, but it can instantiate the following set in the MEASURE-TYPE system in Figure 5.3: [class-property], [animate-size], [inanimate-size], [absolute-child-age], [relative-age], [mass-size], [count-size], as well as [neutral-interpersonal] and [positive-interpersonal] in INTERPERSONAL-TYPE. However, many instances of features from the MEASURE-TYPE system can be filtered out because of restrictions imposed by the modified noun. For example, a reference to an object of the class of children (girl, son, child) sets [absolute-child-age] as the most probable option; alternatively, a reference to a little table leaves the only possibility of [inanimate-size] specification, a reference to money leaves the only possibility of [mass-size] specification, etc. When a pronoun is modified (When I was little), the task is to relate known meaning intentions about the object identified by the pronoun, as well as the reference time used in the clause, to the most probably feature in the set. In the case of a machine translation system, the result of understanding is represented as a list of inquiries with answers ranked in terms of their probability. In the case of translation, we can also assume that the set of meaning intentions is the same for a source text and its translation, though we cannot assume this in general (even for texts written in the same genre), because sets of meaning intentions may differ for sociocultural patterns of uses for different languages, in particular, for interpersonal options.

Under this assumption we can use a uniform set of meaning intentions for guiding generation of texts into other languages. Generation should not necessarily lead to the choice of the same options in the systemic network,
as it was for the source text. There are three levels to accommodate for multilingual differences:

the level of lexicogrammatical realisation:
Of course, different lexical items are chosen for different languages, though this also includes restrictions on the structure of the quality group or realisation at different ranks, for instance, the most natural way to translate *little kiss* in (5.4) into German is by means of *Küsschen*, a word for *kiss* with a diminutive suffix.

the level of lexicogrammatical options:
This involves two possibilities: 1) some options in the network are more or less specific for a language and 2) the same set of options is available for several languages, but different options are selected for a specific language. With respect to the first option, German and Russian lack resources for specific references to the height of animate beings, like the option distinguishing *high* and *tall* in English. This means that the option animate/inanimate in Figure 5.3 is not effective for German and Russian.

As for the second option, the degree of one-to-one correspondence between options is relatively high in the three languages, i.e. size adjectives, like *large, high, long, wide*, regularly correspond to *groß, hoch, lang, breit* and bol’shoj, vysokij, dlinnyj, shirokij, respectively. This also concerns many non-spatial senses, like *high quality, hohes Qualitä*, *vysokoe kachestvo*, *long list, lange Liste, dlinnyj spisok*. For each dimension, say, *deep*, the correspondence covers wide ranges of uses, e.g. *deep appreciation, contempt, cut, mourning, sleep, sigh* are rendered in German and Russian in a similar way: *tiefe Anerkennung, Verachtung, Schnitt, Trauer, Schlaf, Seufzer* vs. *glubokie priznatel’nost’, prezrenie, nadrez, traur, son, vzdoh*. However, there are few cases, in which another dimension is selected. For instance, *deep delight* is rendered in German and Russian in terms of non-directional properties: *große Freude, ogromnoe naslazhdenie*. In such cases, the modified noun (*delight, Freude, naslazhdenie*) controls the expression of intensity by preselecting options in the network independently from the set of meaning intentions.

Some properties expressed by a size adjective in one language can be rendered as a measurable quality without a reference to the size, cf. the example with *high wind* vs. *starker Wind* and *sil’nyj veter* (lit. strong wind). Other examples of this kind are: *low pulse, schwacher Puls, slabyj puls* (lit. weak pulse), *low visibility, geringe Sicht* (small sight), *plohoja vidimost’* (bad sight). Such cases are also modeled by preselection of specific features in various regions of the network (in this case, outside the domain of size specification)
the level of choosers for realising meaning intentions:

The option is least frequent for the three languages. In German and Russian, there are two possible ways for referring to the smaller side of the scale, thus, the REFERENCE-TYPE system in Figure 5.3 should be more delicate. Some adjectives mean really small, e.g. *klein*, *malen’kij*, *uzkij*: they are also regularly used as translation equivalents for *small*. Other adjectives are used for the description of objects that are smaller than their large counterparts, and are below the average or expected value for them; though they do not reach the smaller side of the scale, e.g. *gering*, *mäßig*, *nebol’shoj*, *neshirokij*. When the option is applicable, the choice is made with respect to several criteria:

1. is it important for rhetorical reasons to distance the description from the smaller side of the scale? If yes, choose [below-average].

2. the size of the object in general: is the object inherently larger than a person? This concerns such objects as mountains, boulevards, airports. If yes, choose [below-average].

3. the size of the object in the current situation: is it possible to consider the object or amount in the particular context as smaller than usual? If yes, choose [little].

There is also an interpersonal difference, which distinguishes uses of [below-average] size adjectives from respective [small] adjectives. When *uzkij* (narrow) or *nizkij* (low) are used, they often include a negative characterisation of a value, e.g. *uzkoe mesto* (bottleneck, lit. narrow place), *nizkoe kachestvo* (low quality). On the contrary, *neshirokij* and *nevysokij* are used, when S’s goal (often in advertisements) is to de-emphasize the negative properties of an object being described, e.g. in comparison to its price: *neshirokij kanal sujazi* (narrow communication channel), *skanirovanie s nevysokim razresheniem* (low resolution scanning), *nevysokie tseny* (low prices).

The description proposed above has been tested by development of a small fragment of a multilingual generation system aimed at correct selection of size adjectives in English, German and Russian. The multilingual network has been modeled using KPML (Bateman, 1997) and existing multilingual resources for English, German and Russian (Kruijff et al., 2000). Since no domain model existed for getting meaning intentions from the environment, or a parser for getting them from texts, input specifications were supplied manually in the format of the Sentence Plan Language, SPL (Kasper and
Whitney, 1989). SPLs consist of concepts and relations as well as inquiry-response pairs that are minimally necessary for specification of the semantics of the utterance to be produced.

The following is an SPL example can be used for generation of size adjectives which can be selected depending only on the head noun:

(5.23) (EXAMPLE :NAME NP-WIDE-2
 :GENERATEDFORM "A wide range"
 :LOGICALFORM (R / OBJECT :LEX RANGE
 :PROPERTY-ASCRPTION (Q / QUALITY
 :REFERENCE-TYPE-Q BIG)))

As the SPL and its output in Figure 5.7 show, the only lexical item explicitly specified is range, while the size adjective (as well as the indefinite article) is to be chosen in the lexico-grammatical realisation network. The operators that govern lexical choices are standard for KPML: lexify, classify, inflectify (e.g. for senses with short forms only), cf. (Bateman and Sharoff, 1998).

In the case of spatial senses, the input SPL provides exact information on possible meaning intentions associated with the object:

(5.24) (EXAMPLE :NAME NP-WIDE-1
 :GENERATEDFORM "A wide table"
 :LOGICALFORM (R / OBJECT :LEX TABLE
 :PROPERTY-ASCRPTION (Q / QUALITY
 :REFERENCE-TYPE-Q BIG
 :ANTHROPOCENTRIC-ORIENTATION-Q ACROSS)))

(in a “complete” generation system such answers should be provided by the environment)
Chapter 6

Verbs of motion

6.1 Introduction

This chapter presents a case study in the domain of verbs expressing “away from” motion in English, German and Russian. Table 6.1 lists the verbs used in the study. The selection is based on the aim to cover most frequent uses of most frequent verbs of motion according to representative corpora. Thus, the study is similar to the study of size adjectives: it should mix considerations from form and function. For instance, the consideration of which verb of motion is frequent is based on word frequency lists from respective corpora: the British National Corpus for English, the XLEX corpus for German and the Reference Russian Corpus for Russian. Some information on types of their uses can be taken as senses from dictionaries: Collins-COBUILD for English, Wahrig for German and Ozhegov for Russian. However, very few information on the frequency of senses is provided in dictionaries, for instance, CCED orders the senses according to frequency ranking. Moreover no reliable statistics are available even on the frequency of word forms for phrasal verbs in German (particles and prefixes frequently code the direction of motion).

To achieve wide coverage of the domain, the study used all English and German verbs with frequency above 30 instances per million words (this corresponds approximately to the 3000 most frequent words in a language) and their combinations with most frequent particles and prefixes referring to “away from” motion. Initial types of uses were provided by dictionary senses, which were extended by attempts to code examples of uses in corpora. Corpus-based study of size adjectives was relatively easy because of the rigid structure of the nominal group: an adjective modifies a noun. Thus, we can
restrict our study to the most important left collocates of respective words. However, it is harder to produce and interpret collocation lists for verbs of motion, because the function does not depend on its immediate neighbour. Nevertheless, some important patterns can be revealed.

The lexical semantics of verbs of motion is a popular research topic. See, for instance, (Miller, 1972), (Talmy, 1972), (Levin and Rappaport, 1992) for English; (Eichinger, 1989), (Gerling and Orthen, 1979) for German; and (Apresjan, 2000), (Grenoble, 1986) for Russian. Examples with such verbs are also widely used in general discussions on lexical semantics, for example, (Bierwisch and Schreuder, 1992), (Langacker, 1988). However, most analyses miss the complete range of options which are available for using verbs of motion; in particular, they pay little attention to processes which are typically expressed as a motion, but do not refer to a change in physical location, for example, *to go through a lot of trouble, to advance the claim, to pull out of the project*. Even for physical processes, many studies deal with basically one meaning for a verb and miss other possible uses. For instance, (Miller, 1972) considers a class of traveling verbs and defines *to go* simply as “to travel away from the speaker”, thus, missing many other types of uses, like *It took us an hour to go three miles or I went home at the weekend*. In contrast, the present study shifts the focus from concepts, which allegedly underlie uses of verbs of motion, to resources that are provided by verbs of motion for communicating various types of experience. Thus, the proposed description is aimed at covering all frequent types of uses of motion verbs from Table 6.1. A few exceptions concern idiomatic expressions, like *leaving aside or where it left off*, but they are better viewed as lexical items of their own.

Lexicographic treatments typically include all possible uses of a word. For instance, the verb *leave* has 17 senses in WordNet, 29 senses in CCED, and 31 sense in the Oxford English Dictionary (the senses for the noun, phrasal verbs and idioms were not counted). In the ECD framework, Apresjan (2000) analyzes 19 senses of the Russian verb *vyjti*, which is one of translation equivalents of *leave*.

In some cases, there are errors or omissions that allow a simple correction. For instance, Apresjan considers the following sense of *vyjti*: 6. *vyjti zamuzh*, to get married [of a woman]. This sense is highly idiomatic and in the proper ECD style should be expressed by a separate lexical entry. At the same time, Apresjan (2000) lists the following uses as idiomatic and does not include them into the list of senses: *vyjti na pensiju* (to retire, lit. to leave to pension) and *vyjti v otstavku* (to resign, lit. to leave to a resignation). The list can be easily extended, for instance: *vyjti v zapas* (to
6.1. INTRODUCTION

| English | come off, come out, drive away, drive off, drive out, escape, fly away, fly off, fly out, get away, get off, get out, go, go away, go out, leave, move away, move off, move out, pull out, quit, run away, run off, run out, walk away, walk off, walk out, withdraw |
| German | abfahren, abgehen, abkommen, ablaufen, ausgehen, ausfahren, auskommen, auslaufen, aussteigen, austreten, gehen, sich entfernen, entgehen, entkommen, herausgehen, laufen, verlassen, verkommen, verlaufen, wegfahren, weggehen, weglaufen, wegkommen, wegziehen ????zurückziehen, i.e. cases of translation equivalence leading to differences between models (not away-from, but return) |

Table 6.1: The most frequent verbs of “away-from” motion in English, German and Russian

retire, in the case of military duties, lit. to leave to the reserve), vyjti v rezerv (to be transferred to a reserve, lit. to leave to a reserve). In the ECD style, they should correspond to a separate sense “to leave an institution”.

However, other problems do not allow for an easy solution. There is no way to classify the following uses in terms of the senses in (Apresjan, 2000):

(6.1) ... ne vyxodja iz programmy.
... without the need to leave the application.

(6.2) Pri vzode i vyzode iz polja formy ...
when [the insertion point] enters and exits a form field

The most natural is to apply the sense VYJTI 1.2 “to move outside”, but the complete definition explicitly states: “a vehicle moves into a more open space”. Example (6.1) does not refer to vehicles, but still may be interpreted as a transition from a closed space (an application) into an open space (the operating system). However, even this metaphorical reinterpretation is impossible in (6.2), since the transition is made between two fields. Anyway, the need to use metaphoric interpretations violates the requirement
of explicit and rigorous definitions. In addition, VYJTI 1.2 states that the agent is necessarily a vehicle, but an example for it in (Apresjan, 2000) is Karavan vyshel iz oazisa (the caravan left the oasis), which does not refer to a vehicle and, most probably, even not to a group of vehicles according to its possible ECD definition.

Similar problems occur with more abstract senses of vyjti. For instance, the following use of vyjti from a Russian translation of Lewis Carroll’s Alice in the Wonderland:

\[(6.3) \text{iz } \text{nego by vyshel vesjma neprijatnyj rebenok.} \]

\[\text{from him would go quite unpleasant child-nom} \]

‘it would have made a dreadfully ugly child’

\[\text{can refer to senses 7.1 or 7.2, but it violates the condition for the increased social status in 7.1 and gaining expertise in 7.2.} \]

On the other hand, examples from the corpus show that many uses of the verbs leave and vyjti correspond simultaneously to several senses according to WordNet and (Apresjan, 2000). For instance, the sentence:

\[(6.4) \text{te } \ldots \text{slozhili ix i vyshlo: 44 kopejki.} \]

\[\text{they } \ldots \text{added them and go-impers: 44 kopecks.} \]

they added them up, and the result was: 44 kopecks.

\[\text{may refer to sense 7.3 (production of something) and sense 7.4 (acquisition of a property), though both senses do not fit completely: the definition of 7.3 assumes a production of an object from some raw material, while the definition of 7.4 assumes the possibility for an object to acquire one of several properties and the result came by chance. Vyjti iz boja (to leave the battlefield) is used as an example in 3.3 (to cease to do), but it also corresponds to physical motion in 1.1. Another example of 3.3 from the military domain, vyjti iz okruchenija (to extricate from the encirclement) in addition to physical motion is related to 3.2, i.e. to cease to be in the state of encirclement.} \]

Apresjan (2000) offers a thorough treatment of a single word; we can expect similar problems in a wide-coverage dictionary, which pays much less attention to individual words. For instance, WordNet contains two subsets of senses of leave: “go away from a place” and “move out of, as of a room, a country, a bus, etc.” In many utterances the distinction between them cannot be drawn, so the use should be considered as ambiguous. Also, Sense 9 (leave alone or undisturbed) is claimed to be synonymous to Sense 4
(leave unchanged or unaltered). The relationship between the two senses is indeed evident, but the nature of synonymy between synsets is questionable.

Another group of almost synonymous synsets is Sense 2 (leave behind), Sense 11 (refrain from taking) and Sense 16 (leave behind, forget). They often refer to the same situation: an object is left, when a person has left. The only difference is the emphasis on the deliberate character of the action in \leave_{11} and its inadvertence in \leave_{16}, while \leave_{2} is their common denominator.

In CCED, the three senses are combined into one description (\leave_{4}). Finally, even non-synonymous senses can be used quasi-ambiguously, i.e. can be applicable to a single use, as in:

\begin{align*}
(6.5) & \quad \text{The clues he left did not establish his identity but they reflected his personality} \\
\text{this refers to Sense 2 (leave behind) and Sense 7 (The water left a mark on the silk dress). On the other hand, WordNet (unlike many other dictionaries) does not distinguish between several apparently different uses of leave in Sense 2. Compare (6.6), a person is left behind because of the break of relationship, to (6.7), persons have been left behind because they have been left alone:} \\
(6.6) & \quad \text{I want you to leave your incidental Dick.} \\
(6.7) & \quad \text{... John who then left us, closing the door upon us, with the utmost tact.}
\end{align*}

WordNet 1.6 has a semantic concordance consisting of some texts from the Brown Corpus tagged with respect to their WordNet senses manually. Even though the level of ambiguity reported in the WordNet concordance is low (703 multiple senses for 234136 tagged word forms), the close look at the data shows that the real level of ambiguity is significantly higher. For instance, the verb \textit{leave} in the following sentence:

\begin{align*}
(6.8) & \quad \text{The general, remarkably courteous, explanation has left basic positions unchanged} \\
\text{is tagged in the WordNet concordance only as Sense 3 (cause to be in a specified state), while it is evidently related to Sense 4 (leave unchanged or unaltered).}
\end{align*}
6.2 Basic options for using verbs of motion

The systemic network that covers verbs referring to away-from motion in English, German and Russian is given in Figure 6.1. It includes the following options:

The Type of Motion: physical vs. abstract

(6.9) He left the room vs. He left the job.

The option of abstract motion is considered in Section 6.3 below. In the case of physical motion, there are three possibilities of lexicalisation: [nonspecific-away], a generic reference to motion \((\text{they went to Houston})\), [specific-away], a specific reference to the carrier or conduct of motion \((\text{they flew out to Houston})\) and [cause], an explicit reference to the cause or purpose of motion \((\text{they moved out to Houston})\). Frequent subtypes in the second case are [by-vehicle] \((\text{drive away, wegfahren, uexat’})\), [walking] \((\text{go away, weggehen, idti})\), [running] \((\text{run away, weglauen, ubeshat’})\), [crawling] \((\text{creep away, wekgriechen, vypolzti})\), etc.

There are important multilingual differences between uses of [nonspecific-away] and [by-vehicle] in English, German and Russian. In English, [nonspecific-away] references like \(\text{leave or go (away)}\) are considerably more frequent than [by-vehicle] ones, like \(\text{drive away}\), especially in cases, when the traveler does not drive, but rides in a car, bus or train. In German and Russian, such situations are expressed using forms of \(\text{fahren, exat’}\) (to go by a vehicle), which are obligatory for referring to long-distance motion, which necessarily uses a vehicle:

(6.10) She has flatly refused to go. (to a summer camp)

\(\text{Sie hat es rundweg abgelehnt wegzufahren.}
\)

\(\text{Ona reshitel’no otkalasj exat’.}\)

(6.11) I am leaving tonight.

\(\text{Ich fahre heute abend weg.}
\)

\(\text{Ja uedu segodnja zhe vecherom.}\)

1There are other options that influence lexical meanings of verbs of motion, e.g. activity vs. achievement \(\text{(Levin, 1993)}\). However, they are not specific for verbs of motion, so their choice should be represented in less delicate portions of the network.

2The order of all examples is English, German, Russian, unless it is specified explicitly.
Figure 6.1: The basic options for verbs of motion
As a result, the joint frequency of *drive* and *ride* in English is about 216 instances per million words, while the frequency of the Russian *exat’* (with respective prefixes) reaches 955 instances per million words.\(^3\)

Non-specific references to motion using a vehicle in German and Russian are possible in very few cases, in particular, when a vehicle departs according to a timetable or follows a determined route: *Avtobus ushel v 9.00* (the bus left at 9.00 according to a timetable) vs. *Avtobus uexal v 9.00* (an occasional bus left at 9.00), see (Raxilina, 2000). Finally, motion using a vehicle in German and Russian can be expressed in a generic way for habitual destinations, like home, work, school: *Jetzt gehe ich nach Hause, Sejchas ja uhozhu domoj* (Now I am going home).

In addition to the motion type, the three languages also have several lexicalisation options which are specific with respect to the cause of “away-from” motion:

1. **[entertainment]** – to leave home for entertainment, e.g. *go out, ausgehen, vyjti proguljat’sja; ??+weggehen, ausfliegen*

2. **[relocation]** – to go to live in a different place, e.g. *move away/out, gehen nach X, vyexat’; ??umziehen, wegziehen;*

3. **[home]** – to run away from home, often “because you are unhappy there” (CCED), *to run away, von zu Hause wegzulaufen, sbezhat’/ujti iz doma;*

4. **[displeased]** – some motion verbs used in imperative can refer to expressing the anger of the speaker: *get away, get out of my house; geh weg, komme weg; uhodite, ubirajtes’; ??(verziehe dich, hau ab)*

5. **[end-journey]** – , to leave a vehicle at the end of one’s journey, the language-specific differences are discussed above??: *get out/off, aussteigen, sojti/vyjti;*

6. **[escape]** – to get away from a place of confinement: *escape, entkommen, sbezhat’;*

7. **[error]** – to go in a wrong direction: *sich verlaufen, vom Weg abkommen, also sich verfahren, verfliegen, this is an example of a communicative intention, which is frequently realised by verbs of “away-from”*

---

\(^3\)The frequency data are from the BNC for English and the Russian Reference Corpus for Russian.
6.2. BASIC OPTIONS FOR USING VERBS OF MOTION

motion in one language (German) and by other means in other languages, for instance, by emphasising the result to get lost, zabludit'sja in English and Russian (to wander off or to stray are much less frequent ways to communicate the situation);

8. [delivery] – to be sent (for a package), most typically as a participle in German: the letter went, der abgehende Brief, pisjmo ushlo.

The Direction: [non-directed] vs. [towards] vs. [away] vs. [return]

(6.12) She ran for two hours vs. She entered the room vs. She left the room.

A lexical item may either designate an inherently directed motion (arrive, depart, descend, return) or be indifferent to its direction (roll, slide, walk). In the case of inherently directed motion, it is possible to express the source and the destination, one of which is treated as a reference point, so that the motion is either away from it or towards it (to enter a room vs. to leave a room). The reference point is the selected point for the Path of motion in terms of (Talmy, 1985). Languages also have lexical resources for additional specification of properties of the Path. In the case of physical “away-from” motion, Russian has the following three options for expressing properties of the source and destination:

1. [inner-halo] – X has left a region; this is the unmarked option which is often realised by verbs with the prefix vy-, like vyjti (go away), vyexat’ (drive away), vybezhat’ (run away), vyletet’ (fly away), etc;

2. [outer-halo] – X has left a region, but remains in the proximity to it; the speaker emphasises the fact that X can return back easily, the option is often realised by verbs with the prefix ot- (otojti, ot’exat’, otbezhat’, otletet’);

3. [outer-most] – X has left a region for a remote destination; the speaker emphasises that X is not expected to return back soon, the option is often realised by verbs with the prefix u- (ujti, uexat’, ubezhat’, uletet’).

At the same time, what is considered as [outer-halo] or [outer-most] depends on the current semantic configuration and not solely on properties of the destination. Example (6.13) from the translation of ‘Alice’ by Demurova:

4The names for options follow the analysis of French verbs of motion by (Asher and Sablayrolles, 1997).
(6.13) My ... uzodili na ulitsu i tselyj den’ igrali v klassiki
\[\begin{align*}
\text{we ... went to street and whole day played classics (a game)}^5
\end{align*}\]

shows that the temporal specification makes it possible to consider the destination (the street) as remote, even though it is clear from context that the place is close to the school. If vyxodit’ or othodit’ were used in this situation, this should mean “for a short time”, so this would contradict the whole day activity. Even though, English and German lack regular lexical means to express [outer-halo] and [outer-most], there are several types of uses that are related to the last options. For instance, CCED lists a sense of quit “If you quit a place, you leave it completely and do not go back to it”.

German prefixes: ab-, aus-, heraus-, weg-, gehen/kommen/fahren

**The Focal Point:** source- vs. destination-focused

(6.14) Forty minutes later they left Colombian airspace vs.

(6.15) This year he has left for St. Louis early to acclimatize himself.

This opposition differs from the direction of motion with respect to the reference point. The opposition is typically related to the fact that the speaker can focus on either the source or destination. Unlike the direction of motion, which is more or less fixed for a lexical item, the focal point of the message can be related to specific senses of a lexical item:

(6.16) They wanted to get out of the country [escape] vs.

(6.17) If you get out, you go to places and meet people, usually in order to have a more enjoyable life [entertainment].

### 6.3 The subcase of abstract motion

Physical motion provides a schema for conceptualising various events in terms of motion from one point to another. Analyses of “abstract” motion in (Langacker, 1987) or “fictive” motion in (Talmy, 1999) show some cases when verbs of motion are used for coding events that do not refer to motion in physical space, like *A fence runs along the road*. However, the complete range of such uses is much more diverse: 52 senses in CCED for verbs from

---

5It has no place in the original, because this is a translation of a pun: *I went to the Classics master ... he taught Laughing and Grief.*
6.3. **THE SUBCASE OF ABSTRACT MOTION**

Table 1 refer to physical motion, and 63 senses refer to abstract motion; the same situation holds in German (72 vs. 132)\(^6\) and Russian (35 vs. 77). The set of uses ranges from changing social relationships to spreading goods to stopping an activity to designating small amount of resources left. Several German and Russian verbs etymologically related to motion, e.g. *auskommen* or *izbezhat’*, are used in modern language exclusively for referring to abstract motion. However, all uses of abstract motion allow a relatively compact description.

![Diagram of basic options for abstract "away-from" motion in English, German and Russian]

Figure 6.2: Basic options for abstract "away-from" motion in English, German and Russian

The systemic network shown in Figure 6.2 aims to cover the most frequent options referring to abstract “away-from” motion in English, German and Russian. Five basic types of abstract “away-from” motion can be recognised: [state-transition], [relationship], [ongoing], [entailment] and [spatial-direction], each of which uses a specific interpretation of the physical model of motion from a source to a destination.

\(^6\)The number of senses in German is large partly because of the wide range of uses of German motion verbs and partly because of a very elaborate system of senses used in the German Wahrig dictionary.
CHAPTER 6. VERBS OF MOTION

[state-transition]
This is a very common type of use, which presents a change of X’s state as motion from an initial state to a target state. There are three possibilities:

1. [transfer] – X goes from one state to another: in this case both the source and destination can be in focus. The available lexical choice depends on the nature of the state, e.g. a deliberate topic change: to leave the topic, vom Thema abkommen, ujti ot temy; avoiding punishment: to get off the charge, der Strafe entgehen, izbezhat’ nakazanija, transferring activity: the job went to X, rabota ushla k X. more Russian examples??

2. [cessation] – X ceases to be in a state: only the source can be in focus. Three options are possible: [end-of-state] - a state ends, this also includes the case when X ceases to function properly, e.g. a light goes out, die Lampe geht aus, lampa vyshla iz stroja; [cease-to-exist] - an object or phenomenon disappears, e.g. a stain came out, der Fleck ist herausgegangen, pjatno soshlo; [amount] - a resource is used up, e.g. supplies ran out, vse zapasy ushli, especially including time and time-related amounts in English and German: time/deadline/visa ran out, die Zeit/Frist/Visum ist abgelaufen (the option in Russian is used in idiomatic expressions: vremja ushlo, sroki ushli) zu Ende gehen, aus der Übung kommen, vyxodit’ iz upotreblenija.

3. [commencement] – X enters a new state: either the source or the destination can be in focus, depending on the perspective. The new state can be presented as the destination of “away from” motion, its causes can be presented as the source. Again, three options are possible here: [new-state] - to come/go out on strike, in Streik treten, vyjti na rynok/uroven’ (lit. go out to market/level), [come-into-being] - an object or phenomenon starts to exist or shows itself, the sun came out, die Sonne ist herausgekommen, solntse vyshlo; the truth came out, die Wahrheit ist herausgekommen, Pravda vyshla na poverhnost’; [issue] - an information source appears, the ban/book came out, das Buch ist herausgekommen, zapret/kniga vyshli.

[relationship]
Another very frequent type of using abstract “away from” motion is to designate that a person (X) ends his/her relationships with Y. The following cases are possible:
6.3. THE SUBCASE OF ABSTRACT MOTION

1. [job] – to quit one’s job, to leave one’s office, aus dem Amt gehen, ujti s raboty;

2. [institution] – to get out of the committee, to leave school, den Ausschuss verlassen, von der Schule abgehen, vyjti iz komissii, ujti iz shkoly;

3. [activity] – to pull out of the project, to withdraw from the talks, sich vom Projekt zurückziehen, ujti iz proekta;

4. [person] – to leave a person, sbezhat’ iz semji (lit. run out of one’s family), or The spouse walked out on him, ihm ist seine Frau wegge- laufen, u nego ushla zhena.

[entailment]
A logical relationship between two states of affairs can be present as abstract “away from” motion from one state of affairs to another: follow (from), folgen (daraus), vytekat’. The choice of the option leads to grammatical consequences, namely, like mental processes, it allows for projection:

(6.18) From that it followed logically that he would not be meeting Hildegarde.

(6.19) Vot i vyxodit, chto mne nado ischeznut’ ‘so it turns out (it goes out) that I have to disappear’

German and Russian often use this option to express one’s assumptions:

(6.20) Ich gehe bei meinem Plan davon aus, dass X
V svojom plane ja isxozhu iz togo, chto X ‘In my plan I assume (go out) that X.’

[spatial-direction]
In this case, motion refers to a particular position of an object;

1. [extension] – a static configuration is expressed using a verb of motion, as discussed by Talmy and Langacker, a fence runs along the road, der Zaun geht die Bahn entlang, vdol’ dorogi idet zabor. German and Russian verbs of motion are frequently used to express the direction of extension: der Weg geht nach rechts ab, Doroga uhodit napravo, the road turns (lit. goes out) to the right; ??in eine Spitze ausgehen, Die Brücke geht über den Fluss; Der Tunnel geht unter dem Fluss lang
2. [aperture] – an aperture allows a view/motion towards a certain direction, *Das Fenster geht auf den Garten, Okno vyxodit v sad* (lit. the window goes to the garden, the window faces the garden); no source can be expressed, the destination is obligatory. In English, the option is quite rare: *The right door goes to a ladder that leads upstairs.*

[ongoing] Yet another possibility to use verbs of motion is to refer to an ongoing activity:

1. [trading] – the feature is actively used in German and Russian and concerns uses of specific motion verbs for evaluation of the dissemination of goods, e.g. *Die Waren gehen gut/schlecht ab, Tovary horosho/ploho rashodjatsja* (lit. the goods go out well/bad; ‘they sell well/bad’);

2. [radiation] a psychic quality that can influence other people is presented as radiating out from a person, e.g. *Influence comes from X, Einfluss geht von X aus, Vlijanie ishodit ot X.*

3. [happening] – an event happens (this is most frequently used for evaluation of the process or the resulting state of affairs), e.g. *the interview went well, das Interview ist gut abgelaufen, interju provshlo uspeshno.* No source or destination is possible.

The last option is used in all three languages, but it is especially frequent in German, which has an array of verbs referring to an event in terms of abstract motion: *abgehen, ablaufen, ausgehen, auskommen, gehen, verlaufen, wegkommen.* The differences between their uses depend either on the subjects of an event (who/what “moves”) or on the object of emphasis; in particular, the result or the progress of an event can be emphasised.

The result orientation is realised by *ausgehen* and *wegkommen.* *Ausgehen* emphasises the result of an event, which can end well or badly, like the result of a game or elections:

(6.21) *Wenn aber f"ur Helmut Kohl die Wahl doch schlecht ausgehen sollte, . . .*  
‘Should, however, the elections for Helmut Kohl end (go out) bad,’

It often collocates with *am Ende, zum Schluss* (at the end) or *noch mal* (once more).

*Wegkommen* provides an evaluation of the results of an event from the subjective viewpoint, often when they contradict expectations:
6.3. THE SUBCASE OF ABSTRACT MOTION

(6.22) *Der Postbeamte ist mit dieser Freiheitsstrafe vorerst gut weggekommen.*

‘The post office employee came off with this sentence relatively good.’

It can also be used to evaluate one’s appearance in a description, like news reports or criticisms, often with the intention of distancing the description from the speaker’s opinion:

(6.23) *… dass die Deutschen bei Dostojewski ganz wie bei Balzac oder Chesterton schlecht wegkommen,*

‘that Germans are shown (come out) as bad by Dostoyevsky and similarly by Balzac or Chesterton.’

In contrast to such uses, *ablaufen* and *verlaufen* refer to the progress of a process:

(6.24) *Alle Spiele verliefen fair.*

‘All games ran fair.’

(6.25) *Wir hoffen, dass das Ostergeschaft ebenso gut ablauft wie der Rest der Saison.*

‘we hope that the Easter sale will run out as good as the rest of the season.’

They often collocate with *alles* (all), *das Ganze* (whole), or references to time spans *der Rest* (rest), *das Jahr* (year), etc.

*Ablaufen* can be also applied to presentation media (a music record, a film, etc):

(6.26) *Er versinkt in den Bildern, die wie ein Film in seinem Kopf ablaufen.*

‘He is immersed in pictures, which run in his head as a film.’

The options for using *gehen* in the sense [happening] are between *ausgehen* and *verlaufen*. Like *ausgehen*, *gehen* refers to the result of a process and often collocates with *noch mal* (once more), but can also refer to the progress:

(6.27) *Weil es uns jahrelang zu gut gegangen ist*  
Since it us year-long too well went,…

‘Since the last year went too well for us,…’

*Abgehen* is often used for this purpose in Southern German dialects.
**Auskommen** lies also between process and result interpretations. It can be used to emphasise what can be reached with (*mit*) or without (*ohne*) specific means:

(6.28) *Da der Motor gut mit sechs Litern auskommt.*

‘Since the engine consumes (comes out with) six liters’

(6.29) *Man kann mit 5000 DM für eine funktionstüchtige Kuche gut auskommen.*

‘One can find (come out with) a functional kitchen for 5000 DM.’

It often collocates with specifications of amounts of money, effort or functional abilities required for reaching the result. Similarly, *mit X auskommen* refers to the quality of lasting relationships with X:

(6.30) *Wir sind gut mit einander ausgekommen.*

‘We lived (came out) together well.’

### 6.4 The difference between logic- and communication-centred descriptions

The network shown in Figures 6.1 and 6.2 is able to accommodate some points of the analysis of the Russian verb *vyjti* from (Apresjan, 2000), in so far as it is compatible with a description developed in the logic-centred MTT paradigm. The natural question is: what is the difference between meaning-as-concept and meaning-as-use analyses? After all, the ambiguity between physical vs. abstract motion in *to leave the battle field* (*vyjti iz boja* in Apresjan’s original analysis) is still represented as the ambiguity between features in the network. Terminal nodes in the network can be considered as identical to senses, while intermediate nodes can be considered as groups of senses.

The view “from below” maps individual lexical items to possible combinations of features in the network. The structure of the description from this viewpoint is similar to the distinction between senses available in dictionaries: a set of features often corresponds to a sense. For instance, WordNet lists the following synset:

(6.31) 8. *leave, depart, pull up stakes – (remove oneself from an association with or participation in; "She wants to leave"; "The teenager left home"; "She left her position with the Red Cross"; "He
left the Senate after two terms”; ”after 20 years with the same company, she pulled up stakes”)

which corresponds to the features [job] and [institution] in the network from Figure 6.2. The two features can be treated as a subsense of leave, even though the range of depart in this sense is much narrower than that of leave, as in Unlike some other departing executives lately, Mr Dworkin . . . , while pull up stakes is very rare (there is no instance of this expression in the BNC).

The major positive impact of the proposed description is that it makes it possible to account for diverse uses of verbs of motion in a relatively compact way. For 70 verbs of motion in the three languages, the model covers 325 senses taken from respective monolingual dictionaries. This can be achieved because of the shift in perspective: the proposed network does not reify types of uses corresponding to features as concepts. A “sense” is instantiated when the network is traversed in the context of the current communicative intentions of the speaker (if the network describes the process of natural language production) or in the context of the current attitude of the hearer (if the network describes the process of understanding).

This allows for greater granularity and flexibility in adjusting uses to their context. With respect to granularity, the proposed network describes different resources used for leaving a vehicle at the end of the journey, while WordNet provides a concept (synset) “move out of; as of a room, a country, a bus, etc” which joins an open list of apparently different concepts of leaving. Verbs belonging to the synset (exit, go out, get out, leave) cannot be used in all situations as synonyms:

(6.32) 7He exited the country.

(6.33) *He went out from the bus.

(6.34) 7He got out from the room.

On the other hand, it is clear that the designers of WordNet tried to reduce proliferation of such concepts as LEAVE-COUNTRY, LEAVE-BUS, LEAVE-CAR. This problem does not appear when lexical items are considered as resources, so that respective features (like [end-journey], [public/private] transportation means, etc) are introduced when it is necessary to reflect the difference in the lexical choice.

With respect to flexibility, the possible ambiguity is dynamic. For instance, in the case of vyjti iz boja (to leave the battlefield), one can consider
two possible traversals or prefer one of them depending on the context without the requirement for two separate concepts representing two senses. If a verb of motion is used in a particular context, the network provides the possibility of linking its use via lexical features to communicative intentions underlying the use. Even though the Generative Lexicon model also aims at a greater flexibility of its lexical descriptions, it is unlikely that the fixed inventory of semantic roles provided in the model can achieve the same result in accounting for all most frequent ways of using verbs of motion as described above. At least, all available descriptions of lexical items, as in (Pustejovsky, 1995), cover only a small subset of word uses.

6.5 The preservation of lexical features in translations

The next natural question is: how multilingual is the network presented? The majority of general options for expressing events of different types as motion are strikingly similar in the three languages. Few options that are significantly different are discussed above explicitly. However, the question is whether or not identical situations are expressed by the same set of resources.

An aligned parallel corpus provides the possibility of checking which features of verbs of motion are kept in translations. The study is based on uses of verbs of motion in “Alice in Wonderland”, its three translations into Russian and one into German. The corpus is small and specialised, however, its size allows all uses of verbs of motion in the three languages to be checked. The English original has in total 333 uses of verbs of motion, 182 of which express physical motion, while 130 refer to abstract motion.\(^8\)

The distribution of possible cases for translations is depicted in Table 6.2:

1. the translation of motion is also expressed by motion of the respective type, these are columns P=P or A=A in the table ([physical] is expressed by [physical], [abstract] by [abstract]);

2. the translation refers to a process of another type, the cases here are \([P]\) or \([A]\) = M (material), X (existential) or S (mental, i.e. saying or sensing);

3. a reference to the event is omitted from the translation, P=0 or A=0;

\(^8\)The study is not restricted to verbs of “away from” motion from Table 6.1. The count in Table 6.2 includes all frequent verbs of motion listed in (Miller, 1972) and their translations.
6.5. THE PRESERVATION OF LEXICAL FEATURES IN TRANSLATIONS

<table>
<thead>
<tr>
<th>Texts</th>
<th>Physical motion</th>
<th>Abstract motion</th>
<th>Temporal</th>
<th>Exclamation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P=F</td>
<td>A=A</td>
<td>T=M</td>
<td>E=E</td>
</tr>
<tr>
<td></td>
<td>F=M</td>
<td>A=X</td>
<td>T=0</td>
<td>E=M</td>
</tr>
<tr>
<td></td>
<td>F=X</td>
<td>A=M</td>
<td></td>
<td>E=0</td>
</tr>
<tr>
<td></td>
<td>F=S</td>
<td>A=S</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F=0</td>
<td>A=0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carroll</td>
<td>182</td>
<td>130</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Demurova</td>
<td>132, 21, 7, 6, 16</td>
<td>23, 8, 85, 2, 12</td>
<td>9, 3</td>
<td>8, 0, 1</td>
</tr>
<tr>
<td>Nabokov</td>
<td>130, 21, 12, 4, 15</td>
<td>28, 13, 71, 3, 15</td>
<td>11, 1</td>
<td>8, 0, 1</td>
</tr>
<tr>
<td>Zahoder</td>
<td>119, 23, 13, 0, 27</td>
<td>24, 13, 77, 3, 13</td>
<td>9, 3</td>
<td>8, 1, 0</td>
</tr>
<tr>
<td>Average Ru</td>
<td>127, 21.7, 19.7, 2.3, 19.3, 25.0, 11.0, 77.7, 2.7, 13.8</td>
<td>9.7, 2.3</td>
<td>8, 0.3, 0.7</td>
<td></td>
</tr>
<tr>
<td>Zimmermann</td>
<td>154, 11, 7, 1, 10, 62, 9, 55, 0, 4</td>
<td>11, 1</td>
<td>8, 1, 0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: The distribution of motion types in translations of “Alice in Wonderland”

4. the original use of a verb of motion expresses temporal properties (the English expression *is going to*); it can be expressed differently (T=M) or omitted (T=0);

5. a verb of motion used for an exclamation, e.g. *Come, there’s half my plan done*. This can be translated as an exclamation (E=E), differently (E=M) or omitted (E=0);

The last two options (T and E) are quite infrequent, so we can focus on expressions of [physical] and [abstract] motion. In the corpus, physical motion in English is typically expressed by physical motion (the P=P column) in both German (153 cases, 84%) and Russian (127 cases in average, 70%). Less often verbs of motion are rendered by processes of another type, for example, by other material processes (P=M: 12% in Russian, 6% in German). Compare (6.35) to its German translation:

(6.35) *tell me, please, which way I ought to go from here?*

(6.36) *textit{willst du mir wohl sagen, welchen Weg ich hier nehmen muss?}*

*‘would you tell me which way should I take’*

English motion verbs can also be reported in a more generic way: 6% of references to physical motion are translated as existential processes in Russian, 4% in German. For instance:

(6.37) *After these came the royal children*

(6.38) *Presently the Rabbit came up to the door*
are translated respectively as:

(6.39) Zatem pojvilis’ korolevskie deti  
Then appeared royal children

(6.40) Krolik byl uzhe u dverej  
Rabbit was already at doors

These expressions refer not to motion, but to its result. In this case, the properties of the process, e.g. its destination, are expressed within the circumstantial subtype of relational processes. Relatively often (11% in Russian, 5% in German) a reference to physical motion is simply omitted from the translation.

In short, the feature [physical] is often preserved in translation. Some more delicate features of physical motion are also preserved, when the target language has lexical items with necessary properties and these lexical items can be realised in context. However, a translator may emphasise a specific property of motion. For instance, the sentence (6.41) and its respective translations are described by the following sets of features:

(6.41) was just going to leave the room,  
[directed away source-fg physical nonspecific-away]

(6.42) совсем уж готова уйти из комнаты  
just prepared leave-inf from room  
[directed away source-fg physical nonspecific-away inner-halo]

(6.43) Она уж готова бежать обратно,  
She already prepared run-inf back  
[directed away destination-fg physical specific-away run-away]

(6.44) и уже готова идти,  
and prepared already go-inf  
[non-directed physical nonspecific-manner]

(6.45) und wollte eben das Zimmer verlassen,  
and wanted just the room leave-inf  
[directed away source-fg physical nonspecific-away]

One translation (6.44) uses [non-directed] motion and does not specify the Path. Another translation (6.43) foregrounds the destination and
express the manner of motion explicitly (as running). Other translations (6.42), (6.45) keep the set of features used in the sources. Nevertheless, the general setup of motion is preserved.

Types of references to abstract motion are not only diverse, but also quite frequent. As Table 6.2 shows, there are 151 cases where a verb of motion in ‘Alice’ does not refer to a motion event; this is comparable to 182 references to physical motion. In other registers, like newspapers or science texts, the number of abstract motion uses is even higher than uses of physical motion. With respect to translations, the set of features is less stable: only 19% of abstract motion in English is translated into Russian as motion (the figure for German is larger: 48%). Most typically (60% in Russian vs. 42% in German), another type of process is used:

\[(6.46) \text{Alisa skoro reshila, chto } \ldots\]

Alice soon decided that \ldots

‘Alice soon came to the conclusion that\ldots’

However, the opposite is also true: abstract motion in German and Russian is frequently used in situations where English uses processes of another type. For instance, the example (6.3) is a translation of *it would have made a dreadfully ugly child.*
CHAPTER 6. VERBS OF MOTION
Chapter 7

The Core Lexicon of Emotions

The chapter presents a contrastive corpus-based study of the core lists of words that can indicate emotions in English, German and Russian. First, it defines a methodology for collecting the core lexicons on the basis of corpus evidence and presents the results for the three languages. Then, it presents another systemic network that links words to communicative intentions they can realise. The chapter ends with a more delicate study of words signaling anger in the three languages.

7.1 Introduction

The study of emotions has a long and venerable tradition in psychology. A paper or book that describes the language of emotions often refers back to works of James, Wundt, if not Aristotle. Yet the topic of the lexicon of emotions has been actively discussed recently (at least for English).

In psycholinguistic studies of emotion, we can find two approaches for defining a word list for emotions. It is not surprising that they can be considered as parallel to the distinction between the communication-centered paradigm, which treats words with respect to their use in communication, and the logic-centered paradigm, which considers words as instances of concepts. According to the first approach, as defended by (Storm and Storm, 1987), emotion lists are based on words that are spontaneously produced by subjects in the process of retelling a story or describing emotional states of participants from a picture or a film, cf. also the recent study by Dewaele and Pavlenko (2002). Since lists of the first type come from subjects with
various degrees of language awareness and habits for talking about emotions, the lists are eclectic. Even though researchers may block expressions that refer to emotions only occasionally, like *fat, old, thirsty*, which are treated as low-quality terms by Storm and Storm (1987:809), the majority of words used by subjects for indicating emotional states are kept in their lists. The list of Storms includes, for instance, such labels referring to pleasure as *fantastic, great, super*.

According to the second approach, as defended by (Ortony et al., 1988) or (Johnson-Laird and Oatley, 1989), word lists of emotions should include only clear instances of emotions that conform to several tests, for instance, they refer to internal conditions, are clear cases of emotion-related states, rather than actions, and have affect, as opposed to behavior, as their predominant referential focus. Words in such lists should also pass some linguistic tests, for instance, both “feeling X” and “being X” are emotions; for instance, since *being ignored* does not refer to an emotion, the word *ignored* should not be in the list of emotions (Johnson-Laird and Oatley, 1989:84). In short, lists of the second type are produced by researchers following explicitly articulated principles.

Both approaches intend to collect comprehensive lists defining the semantic field of emotions in English. However, as the result of different attitudes to list composition, respective lists differ in the inventory of concepts, though they typically include a significant intersection between emotional terms (see the comparison below). The situation with data collecting for German and Russian is even worse, as they lack even widely accepted lists of about 500–1000 terms.

The approach of the second type is better aimed at studying the referential structure of the affective lexicon, i.e. the lists are produced for studying concepts related to emotions, but such lists are less suited for studying expressions that describe emotional states in naturally occurring communication. First, such lists are based on emotional concepts, but other ways for realising emotional concepts are missed. For instance, while the concept FURY is listed, the list does not include *furious, furiously* or *infuriate*. Sometimes nouns are favored for methodological reasons, cf. (Zammuner, 1998).

Approaches of the first type are aimed at studying the range of expressions used for referring to emotions. Thus, they are better suited for studying expression of emotions in corpora, but they collect data in artificial settings, in which elicited responses may differ from normal ways for communicating one’s feelings. For instance, in the classification of Storm and (Storm and Storm, 1987), terms referring to desire (*longing, want, yearning*)
are subsumed under the class of negative emotions, because, in the context of their study they imply the absence of something desired, however, uses of words referring desire often occur in the positive context and with positive intentions, for instance,

(7.1) *People wanted to know who this talented designer was.*

Second, it is intuitively clear that some terms are very rare, e.g. *aghast* or *insouciance*, and they do not belong to the lexicon of everyday speech. However, no statistical data concerning the frequency of uses of emotion words are typically provided in psycholinguistic studies. One exception is the old list from (Davitz, 1969).

Third, terms for emotions do not correspond to uses of words. For instance, *emotion/emotional* is treated as the most general term of this class of words in (Johnson-Laird and Oatley, 1989), yet when it is used in texts, it most frequently refers to a *strong* emotion:

(7.2) *...nor was I sorry to break an emotional series that threatened to burden me with...*

Fourth, psycholinguistic studies pay little attention to the inherent polysemy of emotion lexemes. Happiness/sadness is considered as one of the most prototypical emotions (Zammuner, 1998), but uses of the word *sad* are not uniform. For instance, (CCED, 1995), detects various senses of *sad*, such as *sad mood, sad events* (an unsatisfactory or undesirable event or situation), *sad stories* (they cause sadness), *sad old bikers and youngsters* (the expression designates the lack of respect for the participants), as well as fixed expressions like *I'm sad to say.*

The history of studying emotions in lexical semantics also dates back to origins of the field, cf. (Geeraerts, 1988). The topic was actively discussed, in particular, within the framework of Meaning-Text Theory (MTT), e.g. by (Iordanskaja, 1973), (Mel'čuk and Wanner, 1996), the theory of semantic primitives, e.g. by Wierzbicka (1998, 1999), (Harkins and Wierzbicka, 2001), and cognitive linguistics, e.g. by (Averill, 1974), (Kővecses, 1990). In comparison to psycholinguistic studies the range of emotions used in lexical semantics studies is limited: typically a few words are considered,\(^1\) however, the description includes much more details concerning each lexical item.

\(^1\) (Mel'čuk and Wanner, 1996) is an exception: they consider 40 emotion nouns in German, though the study is restricted to nouns and can be hardly comparable to hundreds of words used in psycholinguistic studies.
In addition to the restricted coverage, research in lexical semantics is mostly based on intuitions of researchers and pays little attention to real uses of emotion words. For instance, Apresjan (2000:213) uses the following explication:

(7.3) *Strakh X pered Y; X’s fear of Y = an unpleasant feeling caused in X by Y; this feeling usually occurs when a person perceives or imagines something which in his estimation presents a serious danger to him; his soul feels something akin to what his body experiences when he is cold; his body reacts to this as it reacts to cold; the person experiencing this feeling wishes to become invisible; if the feeling of danger increases in intensity he may lose his self-control and start running or shouting.*

Even though the definition explains some idioms (like поголодеть от страха, to turn cold from fear) and covers some cases of metaphorical uses:

(7.4) *холодная и как бы чешуйчатая сторона моего внешнего молчания наводила, бывало, на Валерию невероятный страх. The cold and scaly quality of my displeased silence, used to frighten Valeria out of her wits.*

it has no relationship to the most of uses of this word:

(7.5) *Сначала, из страха возбудить подозрения, я охотно платил за обе половины. At first, in my dread of arousing suspicion, I would eagerly pay for both sections.*

(7.6) *Осторожней, прошу вас, - закричала Алиса, подскочив со страха. Oh, PLEASE mind what you’re doing! cried Alice, jumping up and down in an agony of terror.*

(7.7) *Дело в том, что я испытывал страх перед автомобилями. You see, I have a fear of cars.*

Also, the definition in (7.3) does not help in explaining the relationship between different words (*dread, fear, frighten, terror*) functioning as translation equivalents of *стрем* in (7.4)-(7.7). What is the semantic core of *стрем*, so that it can cover meanings of these English words? What are different facets of the meaning of *стрем*, so that they allow its translations
in such expressions? Finally, the definition also does not help in distinguishing between different ways to refer to the state of fear in Russian, e.g. ужас, бояться, опасаться, испугаться, etc. Similarly, the study of expressions of anger by (Wierzbicka, 1998) provides definitions for just three words (anger/angry in English, and гнев, сердиться/сердитый in Russian) and does not address many other words that are used to signal anger in English and Russian.

The current chapter presents a study of the basic means used for communicating emotional states in English, German and Russian. The study tries to overcome both the restricted scope of research in lexical semantics and the restricted coverage of lexico-semantic phenomena of research in psycholinguistics. Also, researchers in lexical semantics often make claims about meanings of lexical items on the basis of their linguistic intuition, but the validity of the claims is limited to properties immediately evident to a researcher. However, as examples (7.4)-(7.7) and the study of words referring to anger (Section 7.5 below) show, some hypotheses do not allow an extension to a larger context of language use. The present study does not rely on intuition alone, it is based on evidence for uses of words from representative corpora.

Section 7.2 outlines the methodology for collecting core lexicons of emotions in English, German and Russian on the basis of corpus evidence. Section 7.3 presents the results of collection for the three languages. Then, Section 7.4 explores the basic options available in the three languages for expressing emotions using the core lexicons. The result of the study is stored in a multilingual database of emotion words, which encodes options for their uses and translations. Finally, Section 7.5 is a case study of specific options for using words that signal anger in English, German and Russian.

7.2 The scope and methodology

The two paradigms in lexical semantics identified above roughly correspond to the two approaches to the emotional lexicon in psycholinguistics. The paradigms offer two complementary perspectives. For instance, the aim of my study is similar to the project of Anna Wierzbicka to describe and compare emotions in different cultures, see her multiple works on the topic, e.g. Wierzbicka (1998, 1999). The similarity also concerns the assumptions of the two studies: emotional concepts are not innate, they are specific for a given culture, but their multitude can be described in terms of some basic categories. Thus, the goals of the Natural Semantic Metalanguage used by
Wierzbicka and the multilingual systemic network described below coincide. The difference between the two approaches is once again related to the difference between the logic- and communication-centered perspectives. The focus of Wierzbicka’s studies is on emotion-related culture-specific concepts, which are exemplified by means of respective words, while the focus of my research is on the relationship between words and emotions: the speaker uses words to realise meaning intentions. The exchange of emotional states is also treated as a social phenomenon, involving negotiation of attitudes in the discourse; for a study of social properties of the emotional exchange, see (Martin, 2000).

Another difference concerns the method of research: Wierzbicka’s studies are based on linguistic intuition and information stored in dictionaries, while the present study is based on representative corpora, even though it takes into account linguistic intuition and dictionaries.

The aim of contrastive corpus-based analysis of basic options for talking about emotions English, German and Russian has led to the three tasks:

1. to collect lists of most frequent words in the three languages;
2. to analyze their uses in naturally occurring communication;
3. to compare their uses across languages (the uses are taken from an aligned parallel corpus of literary texts).

The procedure for collecting word lists is aimed at the more or less reliable detection of cases, when emotional states are expressed in a text. This provides a resource for corpus-based investigations, which can use sublists of the structured list for searching instances of expressions referring to emotions in the naturally occurring discourse (written or spoken).

The orientation of the study towards corpus research differs from word lists of emotions collected in psycholinguistics. Any corpus study is based on simple forms. In this case, we look for words that may signal an expression of emotions, not for emotion concepts per se. However, the goal of searching is to study meaning intentions for expressing an emotional state. Respective meaning intentions can be expressed in several ways:

1. directly by lexical means expressing emotional states;
2. indirectly by lexical means using words with emotional implications, for instance, evaluations (fool, great) or words referring to behavioral properties of emotional states (cry, frown);
3. indirectly by grammatical means, e.g. unhappiness in certain contexts can be expressed by *down*, when it functions as an attribute: *The old man sounded really down.*

4. indirectly by pragmatic implications, e.g. “What?” can express an anger, interest, or indifference or be unrelated to emotion.

The last two options cannot be used in corpus-based studies, because detection of emotional states expressed by grammatical means requires deep syntactic and semantic parsing of corpora, while detection of emotional states expressed by pragmatic means requires understanding of texts. Since detection of emotional states by lexical means is based on simple matching of word forms, even the first two methods do not always produce reliable results. For instance, typical names of emotions like *afraid* can be used for purposes that have no relationship to emotions, e.g. *I’m afraid I can’t help you.* However, such uses are relatively infrequent in comparison to emotion-laden ones and can be filtered out, e.g. using a simple pattern: “I am afraid PRONOUN” without any extra syntactic analysis. Also, in German, even detection of lexical items requires shallow syntactic parsing in the case of separable verb prefixes, e.g. *an-tun* (to harm), *auf-regen* (to annoy), *nach-trauern* (to mourn): *du regist mich auf!* (you’re getting on my nerves).

On the other hand, some words may refer to emotions only occasionally. For instance, *flush* and *transport* are listed in (Johnson-Laird and Oatley, 1989); *flush* can refer to a sudden strong feeling, while *transport* to delight. However, if instances of such relatively frequent words were included in the target list, the noise level would be too high, i.e. the majority of uses that could be found in texts, would not refer to emotions. Because words in corpora are often disambiguated with respect to their part of speech (POS), such words as *content* (adjective) or *long* (verb) can be included in the target list. On the contrary, such words as *cry* or *frown* are omitted from the list of (Johnson-Laird and Oatley, 1989), but for the purposes of corpus studies they should be included, because they provide reliable indications of the emotional state of a person, and can be the only indications available in texts, when we search for expressions of emotions on the basis of lexical cues.

Even though evaluations, e.g. *excellent, good, bad,* and personal traits, e.g. *choleric, optimist, sentimental,* are not considered as emotions, they can also be included in the target list because of the following reasons. First, evaluations provide good indications for the speaker’s attitude towards an
object. Second, intentions aimed at evaluation of an object are systematically related to emotions proper: *good to pleasure* and *enjoyment*, e.g. if I say that *something is good* I declare that I *like it*; a *bad/sad/terrible story* is related to the respective mood. Similarly, personal traits are also systematically related to short-term emotional episodes, which often bear the same name, e.g. *anxious*. Finally, such words are spontaneously produced by subjects, whose task was to describe a story, cf. various studies aimed at elicitation of descriptions of emotionally charged situations, e.g. word lists in (Storm and Storm, 1987), (Scherer, 1988), (Dewaele and Pavlenko, 2002).

When we study verbal expressions of emotions in a corpus, two questions should be considered: why the emotional state of a person is at all expressed and what options are used by the speaker for delivering the emotional state by means of linguistic resources available. With respect to the first question, literary texts constitute a specific discourse type, in which emotional words are used to develop a narrative. Corpus-based research can only study explicit statements referring to emotions, but emotions are not expressed via a closed set of words and are often hidden by participants. However, authors of novels use additional means to present the inner state of interlocutors. In the following examples, nothing in the statements themselves suggests the anger of Alice except the comments from the author:

(7.8) ‘*Is that all?’ said Alice, swallowing down her anger as well as she could.*

(7.9) ‘*There’s no such thing!’ Alice was beginning very angrily.*

Even when emotional words are used in speech, as the spoken language part of the Bank of English shows, they often occur when one participant retells a story or describes an event:

(7.10) *we had someone calling us very very angrily saying that you’re saying that we...*

(7.11) *last time I voted out of fear and it was the wrong reason to vote*

Finally, literary texts provide a lot of contextual information, which helps in analysis of meaning intentions of the speaker. For instance, if we analyze expressions of anger, contextual information provided in texts includes reasons of anger, conditions in which it originated and developed, as well as the history of references to the character as being typically anxious or calm, etc.
7.3. COLLECTING THE LIST OF EMOTIONS

As for the second question: what options are used for telling about emotions, the choice of words for telling about emotions in literary texts may differ from the lexicon used for other types of communication, most importantly from expressing emotions in ordinary life in spoken language. The most significant difference is related to the fact that literary texts typically use a richer and more expressive vocabulary and the word choice may be idiosyncratic for a novelist. A careful corpus-based study should also compare expressions of emotions in literary texts and in spoken language. Unfortunately, modern spoken language corpora are not comprehensive: they are small and specialised. For instance, the spoken language section of the BNC constitutes just 10% of the total corpus and largely consists of public speech (recordings of lectures, interviews, meetings) and the corpus of London teenagers, so it is not representative with respect to ways of expressing emotions in ordinary communication in English. What is worse, the German and Russian corpora used in the study practically lack the spoken component (the size of the Russian spoken component is about 200,000 words). Thus, literary texts, which try to reflect spoken communication in natural conditions, provide the best currently possible approximation for studying emotional expressions.

The same argument also applies to the use of translations. Translators make their choices in a way different to choices made by persons in similar situations (very much like authors of literary texts in comparison to naturally occurring communication). Also, a translated text describes situations, which may by foreign or less natural to the culture of the target language. However, aligned corpora provide a rare opportunity to check similarities and differences of the lexicogrammatical resources of different languages with respect to expressing essentially the same emotional state in well-defined contexts. This kind of measurable empirical evidence of word uses across languages is not available in bilingual dictionaries, which list typically decontextualised translation equivalents and are biased in various respects, cf. (Sharoff, 2002), (Wierzbicka, 1999).

7.3 Collecting the list of emotions

7.3.1 Methodology

A corpus-based study of lexicogrammatical means for delivering possible meaning intentions in a domain starts with selecting a list of most important lexical items that can serve as reliable indicators for emotions.

Compilation of the list of English emotion words started with the list of
508 words taken from the study in (Storm and Storm, 1987); the words that rarely refer to emotions were excluded. The decision on the frequency of emotion-related uses was supported by the evidence from the BNC and the subset of the Bank of English included in the electronic version of (CCED, 1995). For instance, pressure was left in the list, unlike small and electric, since uses referring to emotional pressure (under pressure, pressure from/on) are more frequent than references to physical force. Similarly, heart in the two corpora occurs typically as a reference to one’s feelings, e.g. broken/joyous heart; her heart was beating, etc. Another words list for English has been compiled on the basis of the list of 31 basic classes of emotions from (Weigand, 1998). The set of lexical items in the classes has been extended by means of the thesaurus of the electronic version of (CCED, 1995); this resulted in a list of 531 words with high variability over parts of speech.

The combination of the two lists (904 words in total with only 135 words in the intersection) has been further extended with the list of 590 words taken from (Johnson-Laird and Oatley, 1989). This added 269 words to the list (many of them were alternate, e.g. fervent was added to fervour). Finally, some morphological variations were added, if they didn’t enter the list already (mostly verbs and adverbs). The final list of 1191 collected emotion-related words has been checked against the list of 4000 most frequent words in the BNC, the frequency data are from (Kilgarriff, 1996). Lexical items were also filtered with respect to their POS tags, e.g. adjectival uses of long or non-adjectival uses of content were not included. Also, even though both nominal and verbal uses of concern can refer to emotions, the corpus study shows that its verbal uses referring to emotions are quite infrequent. This leaves 253 words for the final list considered below (the words of different parts of speech, e.g. care, are counted separately; participles functioning as attributes, e.g. concerned, are counted separately from verbs).

The list of emotion words in Russian was produced by translating the most frequent emotion words from the English list using ORD (2000) and extending it with words taken from (Jarantsev, 1976). This gives 807 lexical items, which were filtered against the list of the 4000 most frequent words in the Russian Reference Corpus. The final Russian list contains 309 words.

The list of emotion words in German was also compiled from two sources: the list of 128 words from Appendix F in (Scherer, 1988) and translations of words from the English list. The total list comprises 664 words.\(^2\)

\(^2\)The figure differs significantly from English and Russian ones, partly because of the difference in emotional lexicons in the languages, and partly because verbs and participles
7.3. COLLECTING THE LIST OF EMOTIONS

7.3.2 Results

The final lists for English and Russian by themselves allow important observations. First of all, various lists of words referring to emotions include many relatively rare words, which do not belong to the core lexicon of English. Also, the most frequent words in the list do not refer to emotional states, but to emotional goals (want, need, demand) or evaluations (good, great, fine). Even though some of their uses are not related to emotions, e.g. I’m not very good at singing, in many cases evaluations include an emotional component, e.g. It’s so good to hear your voice after all this time.

Also, the lists shed light on the claimed paucity of positive emotions: even though there are fewer words referring to positive emotions in the final lists of all languages considered, the relative frequency of references to positive emotions is significantly higher. For instance, the total frequency of words expressing positive emotions from the English list is 7894 ipm vs. 3222 ipm for negative emotions. If only inner states are included into count (because evaluations are indirect indications for emotions), the distribution still holds: 5322 ipm for positive emotions vs. 1786 ipm for negative ones, cf. the studies on the topic in (Averill, 1980; Zajonc, 1968). The count is based exclusively on the frequency of respective words in the BNC. The calculation cannot take into account contexts of their uses, so it includes cases, when emotion-related words are used for non-emotional purposes (cf. the examples with good and afraid above), and it does not distinguish contexts, in which positive or negative emotions are modulated by negation, doubt or the subjunctive mood. Anyway, the diverting factors are applied to both positive and negative emotions, and the study confirms at least that positive emotions are discussed more frequently.

The comparison of the English and Russian frequency lists shows that both begin with words referring to desire (want, need, wish, хотеть, нужно) and evaluations: positive (good, great, хороший, великий) and negative (bad, плохой), as well as curiosity (interest, интерес). There are some differences among uses of the most frequent words: hope, suffer are on the top of the English list, unlike their Russian translations, while счастье (happiness), радоваться (rejoice), любить (love), удивительный (astonishing), обида (offence), горе (grief, disaster) are significantly more frequent in the Russian list.

The comparison of frequency lists can help in studying the cultural influence on the emotional lexicon. For instance, Wierzbicka (1999:401) points and adverbs derived from them, as well as adjectives and respective adverbs were joined into a single entry for counting German words.
an important feature of the Anglo-Saxon culture to suppress the expression of emotions and contrasts it to openness in the expression of emotions in the Russian culture. The frequency lists of emotion words support the difference: the list of the most frequent emotion words in Russian (309) is larger that the English one (253), even though the original Russian set was smaller. Also, the total frequency of emotion-related words in Russian is 23913 ipm against 17456 ipm in English. Such Russian words as счастье (175 ipm) and радость (126 ipm) are significantly more frequent than expressions referring to happiness in English and refer to an emotion stronger than happiness: they better correspond to joy and elation, which are quite infrequent in the English lexicon.

<table>
<thead>
<tr>
<th></th>
<th>verbs</th>
<th>nouns</th>
<th>adjectives</th>
<th>adverbs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>English:</td>
<td>5893</td>
<td>5497</td>
<td>5760</td>
<td>306</td>
<td>17456</td>
</tr>
<tr>
<td>English:</td>
<td>34%</td>
<td>31%</td>
<td>33%</td>
<td>2%</td>
<td>100%</td>
</tr>
<tr>
<td>Russian:</td>
<td>7271</td>
<td>7755</td>
<td>7025</td>
<td>1862</td>
<td>23913</td>
</tr>
<tr>
<td>Russian:</td>
<td>30%</td>
<td>32%</td>
<td>29%</td>
<td>8%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 7.1: Uses of emotion words with respect to parts of speech

However, the comparative study of the complete emotional lexicon saves from potential errors of overgeneralisation, as for example, another claim by Wierzbicka (1999:398ff) that emotions in Russian are conceptualised as activities, which happen with a person, and are typically realised by verbs, whereas emotions in English are conceptualised as states and are realised by adjectives or nouns. The test is given in Table 7.1, which lists the frequency of English and Russian words referring to emotions both in terms of their frequency in respective corpora and their distribution between parts of speech. No preference for verbs in Russian over English cn be detected in the data. The only significant difference is that adverbs are more frequent in Russian. However, it is based on a difference in syntactic patterns: such English constructions with adjectives as He was afraid/glad/sad are rendered in Russian using adverbs:

\[(7.12) \text{Ему было } \text{страшно } / \text{приятно } / \text{гростно.} \]

He-dat was afraid-adv / glad-adv / sad-adv.
7.4 The framework for uses of emotional words

The systemic network is not aimed at classification of words and not at classification of emotional concepts, but at classification of meaning intentions, which refer to emotional concepts and can be realised in speech by choosing a specific word, so that it describes the experience construed as linguistic meaning (Halliday and Matthiessen, 1999). For instance, sad can refer to both emotional state and to evaluation of an information carrier (story, song, etc), so the two types of uses of the word sad correspond to two different sets of features selected in the network. On the other hand, intentions are systematically related to words that realise them by constraining available lexical options.

Features in the systemic network are based primarily on the properties of emotion-related expressions in texts and only secondarily on psycholinguistic intuitions concerning the emotion in question. Out of this reason, the analysis was based on uses of the words in the corpus as well as on their senses detected by (CCED, 1995). When a choice is described as [neutral], this means that other options in the set of features are irrelevant for choosing a lexical item. For instance, the options of [intense] and [moderate] intensity are not applicable for surprise, as it can be either great or little and can freely combine with any degree modifiers. Thus, the word surprise is marked as neutral, so it can be used for designating various possible intensities. This is different from shock, which is naturally used to realise the intention of [intense] [negative] surprise.

The network discussed below benefits from several other classifications of emotion words, including psycholinguistic (Ortony et al., 1987), (Storm and Storm, 1987) and lexicosemantic studies (Mel’čuk and Wanner, 1996), (Bresson and Dobrovol’skij, 1998). Classifications made from the perspective of lexical semantics pay little attention to proper analysis of emotional states (the only distinction typically made concerns the opposition between positive and negative emotions), while psycholinguistic classifications typically miss properties manifested in lexical co-occurrence, e.g. the intensity or style. Synergy of the two perspectives improves coverage of the semantic field.

Acceptance: positive vs. negative vs. neutral.

This is the typical distinction between positive and negative emotions: good, happy, sympathy (positive) vs. bad, sad, antipathy (negative) vs. relatively few neutral emotion states: interest, surprise (though their expressions in texts most typically refer to positive or negative acceptance). This is probably related to the fact that the expression of emotion is aimed at
negotiating the interpersonal relationship with the hearer, so marking the emotional attitude as positive or negative contributes to evaluation of the inner state, object or event from the speaker's viewpoint.

Mel'čuk and Wanner (1996) reject the distinction between positive and negative emotions as ambiguous and prefer the following opposition: pleasant vs. unpleasant, because in their view Schadenfreude should be classified as a pleasant emotion, even if it has a negative social value, while compassion (Mitleid) is an unpleasant feeling, even if it is socially approved. The classification of Johnson-Laird and Oatley (1989), on the other hand, makes no distinction at all between positive and negative emotions. Their five basic classes include happiness and sadness (the positive and negative counterparts) together with fear, anger and disgust, which positive counterparts (e.g. undaunted, calm or delight) are defined as causes or relations with respect to happiness.

Indeed, the classification is not straightforward, as there are several levels at which positive and negative emotions can be distinguished: psychophysical acceptability, social acceptability, and lexical behavior, e.g. collocations of respective lexical items. If we want to make a distinction between several options (and represent them as features in the systemic network), what is the purpose of doing this? In the case of our network, the distinction between positive and negative emotions is used for delienating and ordering lexical options through which the speaker can realise his/her meaning intentions in the current lexicogrammatical context. Some emotions imply a polarity of states, e.g. happiness vs. sadness. Even though the states at the two opposite poles are not completely symmetrical in terms of the content of emotional states (cf. Tolstoy’s “All happy families resemble one another, but each unhappy family is unhappy in its own way.”), words designating the states form a natural opposition, e.g. excitement or patience vs. depression and anger, so that a set of terms constitutes a semantic field which members share some properties and differ in others. For instance, glad, happy, sad and unhappy share some collocates, such as people (boy, girl, woman, customer), events (affair, circumstances, news, situation), periods of time (childhood, day, year).

The principle for assigning the feature of [positive] or [negative] in the network depends on the answer to the following question: does the speaker intend by using this word or expression to communicate his/her approval or disapproval of the emotional state? For instance, the choice of compassion implies a positive attitude from the viewpoint of meaning intentions, while

---

3See also the psychological treatment of the distinction in Averill (1980:19ff).
Schadenfreude—negative, contrary to the classification used by Mel’čuk and Wanner (1996). This is also confirmed by their definitions in (CCED, 1995): Compassion is a feeling of pity, sympathy, and understanding for someone who is suffering vs. If someone is gloating, they are showing pleasure at their own success or at other people’s failure in an arrogant and unpleasant way.\(^4\) There is no reason to classify the two expressions in the network describing word uses in the opposite way. The word similarity search shows that words which lexical behaviour is similar to compassion are honesty, courage, sympathy, integrity, patience, sensitivity, decency, empathy, dignity, respect, patriotism, kindness, admiration, wisdom (see the description of word-similarity detection methods in Section 3.2). Similarly, it is natural to classify defiance as [negative] for our purposes, because it is a negative evaluation of someone’s behaviour, even though a defiant person does not necessarily “feels something bad” (using Wierzbicka’s terms).

From the viewpoint of uses, the acceptance of situations denoted by some words may differ from their basic senses in dictionaries: feeling is a neutral general term, but it is most frequently used in negative contexts in modern English: feelings of anger, hunger, tiredness; feelings are frequently hurt or offended; I had a strange feeling in the back refers to pain, etc. This does not mean that feeling as a word is annotated as a negative emotion, but its uses are typically annotated as realisations of reference to negative emotional states.

**Intensity of emotion: intense vs. moderate vs. neutral**

All emotions that are expressed are strong enough to take care about them, at least, no lexical items referring specifically to small emotions are available in the languages under consideration. However, names for many emotions can be arranged along a scale: fury, anger, irritation. Apart from differences in the lexical choice, the difference between strong and moderate options is based on their possible lexical collocations (Mel’čuk and Wanner, 1996): [intense] emotions do not co-occur with “mitigators” (e.g. slight), while [moderate] emotions are rarely used with intensifiers, and even being intensified, they do not reach the intensity of [intense] emotions: extreme irritation does not reach the level of fury.

**The cause of emotion: apparent vs. unreasonable vs. neutral**

There are specific lexical means for referring to one’s emotions, which lack an apparent reason, e.g. in English petulant means unreasonably angry; also нутные сюрпризы in Russian refers to one’s worries without a cause, сюрпризы - to frighten without a cause for fear. In the case, when causes

---

\(^4\)Schadenfreude is not listed in CCED.
Figure 7.1: Basic options for expressing emotions
are explicit (either because of lexical semantics of the word denoting emotion or the immediate lexicogrammatical context of the utterance), there are three possible options: the self is responsible (e.g. guilty, shame, sorry), there is an external cause (humiliated, pleased, suffer) or there is a predisposition (shy)\(^5\). Finally, the external cause of an emotion can be attributed to a human being (anger, humiliated) or to an event (suffer). Emotions, which causes are attributed to a human being, can have specific interpersonal options, see below.

**The directionality of emotion: directed-at vs. directionality-neutral**

Directed emotions can be directed either at a person or an object, which is considered as the necessary participant of an emotional state. Some words that can be used in this class have specific grammatical properties, e.g. transitive verbs: enjoy, fear, want, and respective nouns, but also care, be interested, worry.

The first option available for [directed-at] cases concerns the type of the participant to which the state is directed: whether it is specifically human (grateful, threaten, trust) or the participant type is not specified.\(^6\) There can be a grammatical difference between the two options, e.g. the two definitions for care from (CCED, 1995):

\[(7.13) \text{If you care about something, you feel that it is important and are concerned about it.}\]

\[(7.14) \text{If you care for someone, you feel a lot of affection for them.}\]

The second option available for [directed-at] cases concerns the diathesis type: whether the emotional state extends towards its object, i.e. the experiencer of the emotional state is the subject of transitive verbs in the active voice (enjoy, miss, trust), or the emotional state extends towards the subject, i.e. the experiencer of the emotional state is the object of transitive verbs in the active voice (disturb, please, upset). Cf. the difference between wish.verb and wish.noun: the first one is a transitive verb, so it is always [directed-at], but the second one can be general:

\[(7.15) \text{The custom is for people to try and eat 12 grapes as the clock strikes midnight. Those who are successful can make a wish.}\]

\(^5\)The options mirror classes 1.1.1, 1.1.2, 1.1.3 in (Storm and Storm, 1987).

\(^6\)Another option is theoretically possible: directed-at-object, but in the lexical stock under consideration there is no word or use type that can refer to an object and cannot refer to a human being. Care from (7.13) can be used with respect to a person: I’m curious to see if there really is another person in this world you care about.
There is a group of two dimensions that pertain to stylistic properties of the expression:

**Style of expressions: marked vs. unmarked**

The most frequent expressions referring to emotions are stylistically unmarked. The marked options are: [formal], [informal] and [literary]. Good test questions for choosing the options are negative: do you expect to encounter this use in an informal setting? If no, then [formal]; e.g. aspiration, assurance. Do you expect to encounter this use in a formal setting? If no, then [informal]; e.g. mad, sick. Do you expect to encounter this use outside of professionally written texts (such as novels, memoirs, etc)? If no, then [literary]; e.g. anguish, longing. The [literary] feature is often realised by means of idiomatic expressions, e.g. blood runs cold for fear.

Another option potentially possible for this choice is [old-fashioned], but by the very nature of the list of most frequent words, old-fashioned words do not occur in it. Also, when an old-fashioned word is used, most often this occurs within the context of professionally written texts. If it is used by a person, who is not aware of its obsolete status, then the intended meaning which is realised by its use cannot be considered under the option [old-fashioned].

**Person type: personal vs. impersonal**

This choice corresponds to the intention of the speaker to present the emotional state in personal terms (the default choice) or impersonalise it. For instance, desire can be expressed more impersonally by choosing: I need vs. I want. The [impersonal] option often co-occurs with [object-experiencer] for directed processes:

\[(7.16) \text{It pleased him to talk to her vs. He enjoyed talking to her.}\]

Finally, there is a complex network of emotional states presented in Figure 7.2.

**Emotional conditions: inner-state vs. evaluation vs. causation**

The very notion of emotion concerns the inner state of a person, an ‘emoter’; this is the default choice in the network (in this respect, animals and social institutions are treated as persons: they can hope, suffer, demand, etc). However, many words in the resulted lists of emotions can be applied to emotion-laden evaluations (good, beautiful, excellent) or they occur mostly with respect to an emotional state (shake, lover, unexpected). Their uses also share the same set of features (acceptance, intensity, expectation, etc). The classification pertaining to evaluation or causation of emotional states is outside of the domain of emotions proper; for various options in the lexicogrammar of evaluation, see (Martin, 2000).
Figure 7.2: Options for expressing emotional states
**Subtypes of inner states: emotional-state vs. emotional-goal.**

According to the terminology used in (Johnson-Laird and Oatley, 1989), desires are treated as emotional goals; this helps to distinguish them from emotional states proper. Unlike causations and evaluations, they share with emotional states the availability of the subject to which the inner state is ascribed; an inner state can possibly have duration and can be judged with respect to its manifestability.

**Subtypes of emotional states: affective vs. cognitive.**

The subtypes of emotional states differ in the degree of their prototypicality as an emotion. Affective states are the most prototypical emotions like happy, afraid, angry, which are considered as basic emotions in (Johnson-Laird and Oatley, 1989), (Zammuner, 1998). Emotions of the cognitive subtype refer to mental states involving significant amount of mental operations: confident, concentration, interested. Note that even though animals can be happy, afraid, or angry, they can hardly be confident or interested. Ortony et al. (1987) list three components of emotional states: affective, cognitive and behavioral, which are combined in various degrees into classes: affective, cognitive, cognitive-behavioral, affective-cognitive and affective-behavioral.

The class of pure behavioral conditions is left out from their list, because an emotion should have a non-behavioral component. It seems that the classification proposed in Figure 7.2 does not require a separate classification of states, because the subclasses of behavioral states can be described by the combination of two features: [overt-expression] and either [affective] or [cognitive]: cold, laughing, wild (affective-behavioral) vs. care, determination (cognitive-behavioral). If some emotion-related uses are clearly related to behavior and cannot be classified in the [affective] and [cognitive] subtypes, they are considered as causations of emotional states, e.g. relax.

The four basic subclasses of affective states are: [happy-sad], [like-disgust], [courage-fear] and [peace-anger]. They define semantic fields, which are available in the three languages and presumably in all languages, because they are related to universal features of human psychophysiology and the human society. However, lexical items in the fields and modes of their uses are language-specific, cf. the treatment of the semantic field of [peace-anger] in Section 7.5 below.

As noted above, (Johnson-Laird and Oatley, 1989) do not distinguish between positive and negative subtypes of affective states and consider happiness, sadness, fear, anger and disgust as five equally basic emotions. The options of [courage-fear] and [peace-anger] in the network are indeed specific in their treatment of the opposition between positive and negative emotions: they start from calmness (zero level, the normal state) and increase upward.
to arousal caused by something negative (thus, entering the abnormal state). As the result, they lack good antonyms (with the exception of personal traits: brave vs. timid). In the cases of happiness and disgust, the emotions stretch from the positive to the negative end of the scale providing antonyms. The difference between fear and anger: the dominant position of the experiencer is possible for the cause of anger, while it is not possible for the cause of fear.

**Duration of the emotional state: long-term vs. short-term vs. characteristic vs. unmarked-duration**

Some emotional states should be persistent to justify referring to them in a specific way, for instance, distress implies a long-term emotion, while a shock can last only for short time. Long-term emotional states are closer to personal traits [characteristic]; there is inherent polysemy in uses of states and traits: often terms designating an emotional state are used to describe a trait:

(7.17) *The room turned and a carousel of anxious faces undulated round her.*

(7.18) *A friend of mine is a very anxious person.*

Some near synonyms have preferences for state vs. trait readings, for instance, calm vs. patient:

(7.19) *Diane felt very calm and unafraid as she saw him off the next morning*

(7.20) *He was endlessly kind and patient with children.*

The [characteristic] option here is similar to “Frames of mind” in (Ortony et al., 1987).

**Overtness of expression: overt vs. covert vs. neutral**

There can be two options concerning the speaker’s evaluation of one’s behavior with respect to public expression of emotion: an emotional state is explicitly manifested, vs. concealed (in the latter case the words collocate with hide or conceal). The relative frequency of the two options depends on the emotion type. For positive emotions, the overt type is more frequent (the contexts in which happiness or sympathy are described as hidden are relatively rare), while negative emotions are frequently presented as concealed: for instance, according to the collocation sampler of the Bank of English, the verb conceal collocates with contempt, ill, anger stronger than
with delight or excitement (its most frequent positive collocates). The options can concern emotional goals: some goals can be explicitly demanded, while some can be secretly dreamed of.

**Activity: activity vs. neutral**

Some inner states imply an activity on the side of the experiencer: smile, wild, demand. Expressions referring to active states naturally co-occur with imperatives and other forms of requests.

**Control of one’s behavior: controlled vs. lost-control**

In the case of referring to inner states for intense and manifested emotions, a person can be described as either having lost control of his/her behavior (mad, wild) or as controlling it (enthusiasm, demand, hate).

**Interpersonal relationship: dominant vs. submissive vs. neutral**

In the case of states directed at a person or caused by a human agent, linguistic resources can help in distinguishing between the dominant position of the experiencer (care, threaten) and the submissive one (humiliated, weep).

The results of analysis are stored in a multilingual lexical database. The database is represented in XML and encodes:

1. general information on lexical items, e.g. their frequency and morphological properties;
2. most significant collocations;
3. examples of uses of lexical items referring to emotions as well as their translations from an aligned parallel corpus;
4. features instantiated in annotated cases in terms of the network of choices.

It is possible to select groups of lexical items according to specific criteria and consult their uses in context. For instance, if emotions in the narrow sense are interesting for a study, a subset of features for words referring only to emotional states or only to affective emotional states can be output. The search can be selective with respect to several criteria, for instance, lexical items that refer to emotional states resulting in a loss of self-control, or lexical items that designate affective states with obligatory manifestation of the state.

The next section provides a sample comparative study of words referring to anger on the basis of the classification of uses stored in the database.
7.5 Words referring to anger

7.5.1 The structure of the lexical field

Emotions related to anger are in the focus of linguistic studies, see, for instance, (Weigand, 1998), Wierzbicka (1998, 1999). Anger is expressed in many different ways and is related to a multitude of other emotional states, physiological processes or social relationships. Thus, it is impossible to design an exhaustive list of lexical items, which are used to refer to anger. For instance, there is a group of evaluations, which are often used for expressing one’s anger: evil, contempt, despise, rude, or for referring to its absence: kind.adj. Some evaluations are relatively language-specific, for instance, the following German and Russian evaluations are much more frequent that their English counterparts: übel (wicked), негодяй (scoundrel), подлец (villain). Anger is the most probable emotion, when aggressive behavior is described, or when a person scolds or rebukes, either explicitly, e.g. She cursed me, or implicitly, e.g. a use like Shut your gob! Halt die Schnauze! Зачмокни нос! most probably is an expression one’s anger. Weep or burst into tears also refer to a state similar to anger. Finally, there are cases of ambivalence between anger and emotional states, when a mood is related to anger, e.g. anxious, disappointment, regret, upset.

<table>
<thead>
<tr>
<th>English</th>
<th>anger, angry, annoy*, calm*, disturb, frown, frustration, furious*, offence, offend*, offensive*, patient, resentment*, temper*, violence, violent, wild,</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Ärger, ärgerlich*, aufgebracht*, aufregen, beleidigen, Beleidigung, beruhigen, böse, empfindlich*, finster*, Geduld, gelassen, heftig, kränken*, reizen, Ruhe, ruhig, schimpfen, sich ärgern, streiten, verärgern, Wut*, Zorn*.</td>
</tr>
<tr>
<td>Russian</td>
<td>возмутиться, возмущаться, возмущение*, злые, досада, злиться, злоба, злобный*, злость, негодование*, недовольно, недовольство*, обида, обидеть(ся), обидно, обижать(ся), орать, раздражать(ся), раздражение, разозлиться, рассердиться, ругань*, ругать(ся), сердито, сердиться, скандировать*, спокойный, спокойствие, спор*, спорить, твердое, терпение, терпеть, успокоиться, хмуриться*, хмырь*, ярость,</td>
</tr>
</tbody>
</table>

Table 7.2: The most frequent words in the field ‘peace-anger’
The preset study considers the most frequent expressions that relatively reliably refer to the affective state [peace-anger] in the three languages. The words are listed in Table 2 (in addition to the 4000 most frequent words, less frequent words with frequency above 10 ipm were also included, they are marked with an asterisk). Evaluations and anger-related mood descriptions are not included, because their uses most frequently realise other communicative intentions than referring to anger, but words referring to insulting or scolding are included. The lists provide additional confirmation of Wierzbicka’s hypothesis that references to emotions in Russian are significantly more frequent than in English and German: the total frequencies of anger words from Table 7.2 in English, German and Russian are respectively 673, 576 and 1214 ipm.\(^7\)

The nature of the emotional state of anger can be encoded in Wierzbicka’s style explications: “X feels something bad, because Y did something bad” (with some language-specific variations) or, more conventionally: Anger is the strong emotion that you feel when you think that someone has behaved in an unfair, cruel, or unacceptable way (CCED, 1995). Russian provides an impressively wide choice of expressions for referring to one’s anger. Even though, English and German inventories are much smaller, they are not reduced only to anger/Ärger. Among German anger nouns studied in (Melčuk and Wanner, 1996): Ärger, Empörung, Groll, Verärgерung, Verdruss, Wut, Zorn, only Ärger can belong to the core lexicon of the modern German (Wut and Zorn were also included in Table 2, because of their relatively high frequency: 12 and 11 ipm according to COSMAS). On the other hand, adjectives and verbs, such as aufregen (to irritate), böse (angry), sich ärgern (to get angry), are frequently used for referring to the emotional state.

[positive-emotion]

The only word in the core English lexicon referring to the positive acceptance is patient.adj (calm does not belong to the core 4000 words; peace was not included, because its uses are mostly related to international affairs). The Russian list includes сдержанный (self-restrained), спокойный (composed), терпеливый (patient), the German list: Geduld (patience), gelassen (calm), ruhig (peaceful), and some of their derivates. The paucity

---

\(^7\)The smaller number of references to anger is German can be explained by the fact that the German corpora available for the presented research are based on newspaper texts, unlike the BNC and the Russian reference corpus. When a smaller German subcorpus based on literary texts is used, the figures are typically higher, e.g. the frequency of Zorn (wrath) in the Mannheimer corpus (2,54 MW) is 34.25 ipm, while its frequency in the total written-text part of the COSMAS corpus (374 MW) is 10.87 ipm.
7.5. WORDS REFERRING TO ANGER

of positive emotions in the lists is explainable by the nature of the [peace-anger] scale: it starts from the normal undisturbed state as the base for positive emotions, but the normal state is rarely mentioned and is hardly counted as an emotion. The positive reference occurs as an observation in situations, when the subject keeps composure and does not react in spite of someone else’s “unfair, cruel, or unacceptable” behavior. The positive state is often expressed by negating the negative reference, e.g. *He did not get angry, Er hat keinen Ärger bekommen, он не рассердился.* In the cultures under consideration, the zero level of anger in such situations is typically considered as positive. The zero level of the state explains also, why no intensity modulation is used in the case of [positive-emotion] (at least, in the languages under consideration).

**[negative-emotion]**

The set of lexical items referring to non-zero anger can be divided into three groups:

1. expressions referring to the state directly corresponding to anger;
2. expressions referring to “submissive” anger;
3. expressions referring to a verbal manifestation of anger.

**Expressions referring to the state directly corresponding to anger**

It easy to detect three groups with respect to the intensity of emotions:

**[neutral]:**

English: anger, angry, temper*, resentment*,
German: Ärger, ärgerlich*, sich ärgern, aufgebracht*, böse, finster*, verärgern,
Russian: возмущаться, возмущаться, возмущение*, злиться, разозлиться,

**[moderate]:**

English: annoy*, disturb, frown,
German: aufregen, reizen,
Russian: досада, недовольно, недовольство*, раздражать(ся), раздражение,

**[intense]:**

English: furious*, violence, violent, wild (for no apparent reason)
German: heftig, Wut*, Zorn*
Russian: гнев, негодование*, злость, злобно*, злобный*, злость, ярость
Other ways for expressing [moderate] anger, e.g. *irritation*, are not frequent in English and German according to the BNC and COSMAS. Frequent words referring to intense emotions in English and German most typically have other types of uses, however, according to corpus studies they provide a reliable heuristics for detecting states of anger in corpora. The attribution of certain words to a particular position in the scale of moderate-neutral-intense is fuzzy, since it depends on the context of their uses. Also, the distinction between degrees of anger is fuzzy from the viewpoint of translation equivalence: a translator is relatively free to specify the degree (or, rather, free to weight the intended rhetorical impact on the reader). For instance, **раздражение** (irritation) can be presented as equal to anger:

(7.21) *I was angry, disappointed and bored,*

(7.22) *Я чувствовал раздражение, разочарование и скучу*  

(7.23) *The Queen turned angrily away from him.*  

(7.24) *Королева с раздражением отвернулась от него.*  

(7.25) *She had been annoyed by Lo’s liking me.*  

(7.26) *Ее злило, что я девочке нравился.*  

(7.27) *Alice felt a little irritated at the Caterpillar’s making such VERY short remarks.*  

(7.28) *Алиса немного рассердилась – уж очень неприятно говорила с ней Гусеница.*  

The corpus-based study of the lexical field also elucidates some claims made from the perspective of linguistic intuition. For instance, Wierzbicka (1998:20) claims that “Russian does not have a noun corresponding in meaning to the English noun *anger* (because *smeя*≠*anger*)”. However, as the lists above show, there are many Russian nouns in the field and they are often used (according to the aligned corpus) as translation equivalents for *anger*, e.g. *ярость* (most often as a part of the following pattern: **прийти в ярость**, lit. to come into anger), *негодование*, *раздражение*, *злоба*, *злость*:

(7.29) *swallowing down her anger as well as she could.*  

с трудом сдерживая негодование. (lit. with efforts restraining anger)
What is more, *anger* and *gnev* are often used as perfect translation equivalents:

> she stared at me with those unforgettable eyes where cold anger and hot tears struggled.  

> она пристально смотрела на меня этими своими незабвёнными глазами, в которых ледяной гнев боролся с горячей слезой.

> she gave a little scream, half of fright and half of anger, and tried to beat them off.  

> Она вскрикнула — полуспуганно, полугневно, — принялась от них отбиваться ... 

This does not mean that *gnev* is a translation equivalent for *anger* in any context, but the aligned corpus gives evidence that *gnev* is frequently used for realising the same intention as *anger* in English. 9 times out of 15 in the aligned corpus *anger* is translated by *gnev*, though the sample is too small to be representative. For instance, another translation of (7.33) uses another anger-related word:

> Алиса вскрикнула — полуспуганно, полусердито, — стала от них отбиваться ...

The structure of the Russian lexicon of direct references to anger is based on three types of terms: *взмутиться, сердиться, злиться*, which uses mostly differ with respect to the presentation of the cause of anger: the cause of the emotional state designated by *злиться* is typically unjustified (from the viewpoint of the speaker), but *взмущение* (as well as *gnev*) realises the meaning intention referring a well-founded state of anger, while *сердиться* is neutral in this respect. Also, even though *злоба* is morphologically related to *злиться*, it is typically used to refer to more intense emotions in comparison to *злиться*. 
Expressions referring to “submissive” anger

The anger states described above imply an active response against the cause of anger: “I want to do something in response, and I am not dominated by the causer”. Another type of response is also possible: the subject is not inclined to fight against the causer. The subject in this case is submissive with respect to the cause of anger.

- English: offence, offensive*, offend*,
- German: beleidigen, Belidigung, *empfindlich, *kränken,
- Russian: обидя, обидеть(ся), обида, обижаться(ся)

This is an example of a state, which is foregrounded in some cultures and downplayed in others; it is rarely discussed in the studies on anger in the English-based tradition, because its lexical manifestations are not typical in English: offence is a frequent noun, but it often refers to legal notions, while offend is relatively rare (9.87 ipm in the BNC, to be offended – 5.03); it can also be used in the sense of generic anger:

(7.36) “It’s a pun!” the King added in an offended tone, and everybody laughed.

Unlike offended, German and Russian lexical resources referring to the “submissive” state cannot assume that the offended person has a dominant position. Out of this reason, the example (7.36) cannot be translated as beleidigt or обижено, so translators used words referring to anger of the King:

(7.37) Es ist ja ‘n Witz! fügte der König in ärgerlichem Tone hinzu; sogleich lachte Jedermann.

(7.38) - Это каламбур! - закричал сердито Король. И все засмеялись.

Expressions referring to a verbal manifestation of anger

- English: -
- German: schimpfen, streiten
- Russian: ругать*, ругаться(ся), скандальть*, ссоря*, ссориться

In English and German there are also many options for referring to verbal manifestations of one’s anger, for instance, curse, quarrel, scold, swear or brüllen, fluchen, grollen, sich zanken, however, they are rarely used in the written speech according to the BNC and COSMAS. There are also relatively frequent lexical options which are homonymous with other uses, for instance,
argue (disagree in an angry way), row (an angry argument) or Krach haben (to have a quarrel), though the relative frequency of their anger-related uses is not known. In contrast to this, direct lexical references to verbal expressions of anger are routine in Russian. Also, Russian translations often render generic verbs, like say or call, used in English source texts by means of anger-related ones:

(7.39) ‘I wish I hadn’t cried so much!’ said Alice

(7.40) Зачем ты только столько рыдала, дурочка! — ругала себя Алиса. lit. ‘cursed herself Alice’

(7.41) Lo was enraged by all this—called me a lousy crook and worse

(7.42) Лолита приходила в ярость, ругала меня паршивым жуликом и еще худшими словесами

7.5.2 A case study of anger references in a story

The classification helps in making a detailed analysis of references to anger states in texts and their translations. For instance, we can consider the anger-related behavior of the Queen from “Alice in Wonderland” and its translations. The Queen is explicitly presented as a character prone to bursts of anger (the short story contains 19 references to Queen’s anger). Most frequently, anger outbursts of the Queen are expressed 1) by references to the way she says things, 2) by references to qualities of her actions or inner states, and 3) by observable symptoms indicating her inner states. There is no instance of “submissive” anger.

References to verbal expressions of anger are the most frequent: there are 13 instances (this is explainable, because the story largely consists of dialogs), e.g.

(7.43) when the Queen jumped up and bawled out...

(7.44) “Leave off that!” screamed the Queen.

(7.45) as she heard the Queen’s voice in the distance, screaming with passion

There are 4 references to qualities of actions or inner states, typically by means of adjuncts, while the Queen is the grammatical subject of actions, e.g.
(7.46) ...said the Queen, tossing her head impatiently

(7.47) in a very short time the Queen was in a furious passion

The majority of expressions of Queen’s anger (both verbal and non-verbal) are not directed from the viewpoint of the immediate lexicogrammatical environment. Since the directedness of anger specifies its cause, Queen’s anger is presented as unfounded.

There also two references to observable bodily symptoms:

(7.48) The Queen turned crimson with fury

(7.49) “Hold your tongue!” said the Queen, turning purple.

Such references were not studied among the most frequent expressions of anger above, however, according to the BNC colour words are frequently used to refer to the expression of one’s emotions, especially turn/go crimson; be crimson with anger/emarrassment/laughter/rage, but also turn pale, whiten. The same applies to Russian and German: побагроветь (redden) is used in the Russian translations and purpurrot werden (become crimson) in German.

Translations follow the original in their references to Queen’s behavior, though the lexical choice varies in different Russian translations. The most volatile property of translations is the intensity of emotions, e.g. in translations of (7.21–7.27) an angry state can be presented on the range from wrath to irritation depending on the context and rhetorical goals of the translator. What is more important is that translators keep stylistic properties, i.e. they choose expressions according to the classification of formal, informal or literary options. This is yet another confirmation that words are closer related to communicative/rhetorical intentions than to “objective” properties of emotional states:

(7.50) as she heard the Queen ” s voice in the distance , screaming with passion

(7.51) Алиса услыхала , как Королева что - то громко кричит вдалеке

(7.52) Издали доносились голос Королевы , которая орала в яростном испущениин

(7.53) До нее еще издали донесся голос , вернее , вопль разъяренной Королевы .
7.5. WORDS REFERRING TO ANGER

(7.54) da sie in der Entfernung die Stimme der Königin hörte, die vor Wuth außer sich war.

As for multilingual differences, there are differences between translation equivalents: злоба, as the most frequent Russian translation equivalent for angry, also is used in the sense evil. Обижа́ться corresponds to ‘be offended’, but it can be often considered as the state of anger of moderate intensity, because it is frequently used in this function and can be translated as irritation, and vice versa, angrily is often translated as обиженно in the story:

(7.55) ‘Then it wasn’t very civil of you to offer it,’ said Alice angrily.

(7.56) Зачем же предлагать? Это не очень-то вежливо! — обиженно сказала Алиса.

Another feature of German and Russian basic lexicons is relative frequency of words referring to scolding. Offence and its derivatives are not as frequent in the English lexicon in comparison to its German and Russian counterparts, however, it is frequently used by Lewis Carroll:

(7.57) in a whisper, half afraid that it would be offended again.

(7.58) Последнюю фразу она произнесла шепотом, боясь, как бы не обидеть Мышь снова.

(7.59) fügte sie leise hinzu, um nur das niedliche Thierchen nicht wieder böse zu machen.

(7.60) over, and she felt certain it must be really offended.

(7.61) Шерстка у Мышки стала дыбом. Алиса поняла, что оскорбила ее до глубины души.

(7.62) denn diesmal sträubte sich das ganze Fell der armen Maus, und Alice dachte, sie müsste sicherlich sehr beleidigt sein.

(7.63) ‘But you’re so easily offended, you know!'

(7.64) Просто вы все время обижаетесь.

(7.65) “Aber du bist so sehr empfindlich, du!”

(7.66) I wish the creatures wouldn’t be so easily offended!
CHAPTER 7. THE CORE LEXICON OF EMOTIONS

(7.67) A про себя подумала: „До чего они тут все обидчивые!“

(7.68) Bei sich dachte sie: „Ich wünschte, alle diese Geschöpfe nahmen nicht Alles gleich übel.“

(7.69) tone, ‘I ’m afraid I ’ve offended it again!

(7.70) – грустно промолвила Алиса. – По-моему, я ее опять обидела!

(7.71) sagte Alice in reumütigem Tone. “Ich fürchte, ich habe ihr wieder weh gethan!”

(7.72) had nothing yet, ‘Alice replied in an offended tone, ‘so I can’t

(7.73) – Еще? – переспросила Алиса с обидой. – Я пока ничего не пила.

(7.74) ich habe noch keinen gehabt, ‘antwortete Alice etwas empfindlich, ‘also kann ich nicht noch mehr trinken.

7.6 Conclusions

The research reported in the paper has two goals. The first one is to define the core lexicon of words that are frequently used to refer to emotions in English, German and Russian. The goal has been achieved by compiling lists of words that can provide a good indication of emotion-related expressions and by filtering the lists of emotion words against respective frequency lists to select most important (i.e. most frequent) words. The second goal of the research is a network of features that can provide a sensible explanation of any emotion-related use of any word from the core lexicon (and, if possible, beyond it). The network also provides the basis for comparing uses across languages. Expressions with similar intentions receive similar sets of features in the network. The resulted lists of the most frequent words that denote emotions in the three languages under study are stored in a multilingual database. The database can be used for various purposes, for instance, for selecting specific words in psycholinguistic studies, for studying translation equivalence in uses of emotion words, for second language education, etc, because it allows to select groups of emotion words and to study contexts of their uses.
Chapter 8

Conclusions

The aim of the study reported in the book is to build a model for the contrastive description of lexical items from the functional perspective. There are two contributions of the study: the theoretical and the practical one.

The theoretical contribution concerns the comparison of two models: one for describing lexical meanings in a formal framework, which assumes that words are pointers to senses, and the other one in a functional framework, which assumes that words are resources and are used to realise communicative intentions of the speaker. The corpus-based study of uses of size adjectives and verbs of motion illustrates that even though lexicographic research aimed at enumerating senses of lexical items provides an invaluable resource for describing typical meanings, the theoretical model of word uses cannot be based on the notion of a dictionary-like mapping between words and senses, because real uses of words in texts often violate the necessary and sufficient conditions for treating them as instances of specific concepts: uses often refer to slightly different concepts or refer to several concepts simultaneously.

At the same time, the communication-oriented model can achieve a more concise and robust description of basic types of uses for sets of lexical items in several languages, by using the systemic network for relating communicative intentions to lexical items. The objective of the case studies follows Halliday’s objective to construct a grammar “that would make it possible to say sensible and useful things about any text, spoken or written, in modern English” (Halliday, 1985:xv). Similarly, my objective in the study was to say sensible and useful things about the most frequent uses of words from some lexical classes in English, German and Russian, i.e. to show what resources they provide for communicating various types of situations in the
three languages.

Leaf nodes in the network of lexical features may look similar to senses in WordNet or ECD. However, the crucial difference of the lexical systemic network compared to the dictionary-based approach is that sets of features do not constitute concepts. WordNet and ECD assume that words mean something, whereas the proposed model assumes that words are used to realise realise some functions. Yet, the realisation is not arbitrary, as the network of features specifies the potential for using words. The potential is instantiated by a traversal of the network in context, but the traversal does not lead to reification of a concept to which a word points. This means that the meaning of a lexical item in the context depends on its contribution to the ongoing discourse. The systemic network can also address the issue of similarities and differences in uses of words in the three languages without the need to map between language-specific concepts.

The description is not only pure discursive, it is stored in a lexicographical database covering about 750 headwords in the three languages in total (Sharoff, 2004a). This opens several lines of research. First, the database helps in other descriptive studies as well as in development of usage-oriented dictionaries. A database of this type gives responses to queries that concern the set of lexical items possible in a given context in a given language. For instance, the database lists possibilities to refer to ‘escape’ in English as escape, flee, get out, run away, as in

(8.1) A prisoner has escaped from a jail in northern England
(8.2) They probably wanted to get out of the country
(8.3) Palestinian boys run away from an Israeli tank in Jenin. (Reuters photo)

Then we can compare them to the possibilities that exist in Russian, such as sbezhat’ and ubezhat’. Similarly, the set of possibilities of referring to leaving one’s job in English and Russian includes, respectively, go, leave, quit and ujti, vyletet’, pokivut’, even though in the case of non-congruent realisations of concepts in terms of motion or size, the database currently does not include other expressions possible in the field, such as retire or uvolit’sja for leaving one’s job.

On the contrary, the database of emotions starts from the other end by classifying all possible references to emotional states. For instance, it can show explicit expressions referring to the sudden short-term bursts of strong anger, such as livid, furious, loose one’s temper, seething with anger, or non-congruent expressions (often related to the [informal] feature), such as hit
the roof, go nuts, together with other expressions possible in Russian with the same set of features, such as russvorepet’, byt’ v jarosti, zverstvotat’, or informal noncongruent expressions such as sletet’ s katushek or . Further specialisation of queries is possible with respect to expressions of [dominant] or [submissive] anger. This type of relationship can be useful for research in the domain, as well as in teaching foreign languages.

Some information of this type is available in thesauri, for instance, Random House Webster’s Dictionary lists for one sense of escape:

break free, break loose, get away, make a getaway, make off, slip away, run away, flee, skip, bolt, cut and run, abscond, steal off;

Slang fly the coop; avoid capture

which are appropriate for the [escape] feature, but it does not provide information on contexts in which the words are used to indicate “away-from” motion to escape from a dangerous situation. After all, the following example does refer to a dangerous situation, but fails to refer to a motion away from it:¹

(8.4) The only sure method of avoiding capture, she tells me, is to walk backwards at all times, flicking every particle of dust from every footprint . . .

The reason for the discrepancy is once more related to the difference between the concept- and function-based models. thesauri are inclined to classify concepts, not giving functions for using lexical items (in the same way as dictionaries and even more). Also they typically provide only a straightforward tree of concepts, while the systemic network results in a set of interrelated options, so that there is a variation, for instance, for the level of formality, alterations in the degree and duration of emotional states, or the diathesis (X infuriates me vs. I’m furious because of X), etc. Finally, the database allows the comparison of functions across languages. In this sense, they are similar to bilingual dictionaries, but once again do not indicate functions of uses and suggest contexts for appropriateness translation equivalents. They also offer a limited set of translation equivalents, see the only translation of anger available in the Oxford Russian Dictionary and the variety of options discussed in Section 7.5.

The second line of research possible with the databases developed concerns their use in computational applications such as multilingual generation

¹The same applies to the majority of uses of make off, skip, steal off, etc.
and machine translation. The association between features of the networks and lexical items is typically many-to-many, reflecting the inherent polysemy of lexical items (there are many functions that can be realised by a lexical item) and the variety of expressions possible in a given communicative situation (there can be many lexical items available for realisation of a feature in a given context). The fact that the systems of functions and lexical patterns that realise them behave differently in different languages leads to problems in tuning existing dictionaries in multilingual generation and machine translation, because lexical items in a computational dictionary typically offer direct translation equivalents, thus reducing the variety of translations suggested in a bilingual dictionary (which itself does not cover the variety of translations possible in a context) to a set of very few (typically one) equivalents. For instance, Systran, one of the leading machine-translation systems (SYSTRAN, 2004), gives the following German (8.5, 8.8, 8.11) and Russian (8.6, 8.9, 8.12) translations for the original texts in English (8.7, 8.10, 8.13):

(8.5) Ich wollte sie weg irgendwo erhalten zu sicherem.  
I wanted her away somewhere receive to safe

(8.6) Ja zotel poluchit' ejo otsutstvujusche k gde-to bezopasnoj.  
I wanted receive her absent to somewhere safe

(8.7) I wanted to get her away to somewhere safe.

(8.8) Ich glaubte dem unermesslichen Stolz und der tiefen Freude.  
I beleived the huge pride and the deep delight.

(8.9) Ja chuvstoval bol'shuju gordost' i глубокое naslazhdenie.  
I felt large pride and deep delight.

(8.10) I felt the immense pride and deep delight.

(8.11) Er ging über es verrückt, bevor es ...  
He went over it crazy, before it ...

(8.12) On poshol shal'nym nad im prezhde chem ono ...  
On went crazy over him before that it ...

(8.13) He went crazy over it before it had broken into leaf.

In the first case, the system pays no attention to the fact that get is used in the sense of motion: it uses erhalten and poluchit' (to receive) as the translation equivalents of get. However, even if the phrasal verb was
detected and a verb of motion was used in translation, the system would still have a problem with rendering the sense of escaping from a (dangerous) place (probably as run away, not the neutral go away). In the second case, the system pays no attention to possible collocations between translation equivalents of immense and deep in the target languages. In Russian gordost’ (pride) typically collocates with velikij (great) or is left without a qualification (it is assumed to be great), even though bol’shaja gordost’ occurs in the corpus (4 times per 50 million words). In the case of naslazhdenie (delight), it cannot be used with glubokij (deep), deep delight is typically translated as bol’shoe/ogromnoe naslazhdenie (big/huge delight). In German the situation is reverse, whereas Freude (delight) can be combined with tief (even though, it is not its standard collocate), Stolz cannot collocate with unermesslich (huge), instead, when the “size” of pride is discussed, it collocates with ganz (whole), voll (full), groß (large), or gewiss (certain, in the sense of size). In the third case (8.13), the machine translation system does not detect the reference to an emotional state, which is not rendered by a verb of motion in either of the target languages, so translations offered are meaningless.

Since the mechanism of systemic networks provides the description of basic classes of functions in their relationship to their syntagmatic realisations, including the immediate context, it can be used to solve the problem by associating the lexical item in the context of other words with the function it, most probably, realises in the sentence. This will be [escape] in (8.7), [mass-measurement] in (8.10), or [emotion] in (8.13). Then the generation component will select the appropriate realisation of this function in a target language.

The situation is even more straightforward in the case of multilingual generation, as there is no need to discover functions realised by a lexical item. The process of generation starts with the set of communicative intentions, which are ‘worded’ in terms of lexicogrammatical functions. Then, the multilingual lexicogrammar realises the functions by specific lexical items and grammatical constructions specific to a target language, cf. the suggestions on the principles of multilingual grammatical networks in (Bateman et al., 1999, 2000). The only experiment in the multilingual generation technology done with the database so far concerned lexicalisation of size adjectives in nominal groups for English, German and Russian (see Section 5.6).
Bibliography


Hendrik Feddes. XLEX suite of tools for linguistic data processing, 2002. URL http://xlex.uni-muenster.de/.


Irina Kononenko and Serge Sharoff. Understanding short texts with integration of knowledge representation methods. In D. Björner, M. Broy,


from the French by Robert Czerny with Kathleen McLaughlin and John Costello.


