Structuring the Lexicon

Cognitive Linguistics Research 43

De Gruyter Mouton

Structuring the Lexicon

A Clustered Model for Near-Synonymy

by
Dagmar Divjak

De Gruyter Mouton

To my grandmother
Marie-Louise Vanwelckenhuysen

Preface

This book has a long history. Many of the ideas and methods presented emerged while I was working on my PhD dissertation, funded by a four year research assistantship from the Research Foundation – Flanders, Belgium (2000–2004). Yet "going linguistics" was and still is not an obvious choice for a Belgian Slavicist and it would not have been an option for me either if it weren't for the people who took me under their wings during those early years.

In this respect, thanks are due to the late Karel van den Eynde, who got me hooked on linguistics (but would most certainly sincerely dislike this book), for the many Wednesday afternoons that I spent behind his desk in Blanden, breaking my head over patterns from Lingala, Çokwe or any other language Karel knew and I did not. To Laura Janda, my "adoptive advisor", I am indebted beyond my ability to repay, for granting me cognitive linguistic asylum in Chapel Hill and advising on every single aspect of a budding linguist's life. Warm thanks also go to Maarten Lemmens, for luring me into cognitive corpus-linguistics, and to Stefan Th. Gries, who tuned my brain to "statspeak" and explored many aspects of behavioral profiling with me. I would also like to express my gratitude to Jurij D. Apresjan and his collaborators for letting me participate in their working sessions on the Dictionary of Synonyms and to Tatjana E. Janko and Ekaterina V. Rachilina for the many, many hours they spent discussing my data and analyses with me. The in all nine months I spent in Moscow were made possible thanks to a researcher exchange scheme between the Flemish Community and the Russian Federation.

As corpus data for Russian became easier to come by and statistical procedures less daunting to run, analyses improved and theoretical implications emerged more clearly. This turned a dissertation that used insights from cognitive linguistics to make sense of the data, into a book that aims to make a contribution to cognitive linguistic theory, illustrating its main claims with Russian data. I hope this monograph will enrich understanding of established aspects of the cognitive model of language and serve as catalyst for their further development and refinement.

Crucial for this development was not only time, but also location. With the financial support of the FWO-Vlaanderen and the BAEF, this book left Belgium in 2004 and has since lived on desktops in offices at the UNC-Chapel Hill (USA), Stockholms Universitet (Sweden), and the University of Sheffield (UK). It goes without saying that I have benefited enormously from the knowledge of

colleagues-linguists at those institutions. The same holds for comments made by audiences at conferences where I presented parts of my work. Their contributions are gratefully acknowledged at the appropriate junctures throughout this book, which by no means implies their approving of the final product.

One reference that is not mentioned in the main text is Oesten Dahl's sublime manual for PhD students *Hur man undviker att disputera* ('How to avoid graduation'). Yet it should be mentioned somewhere, since work on this book has provided rather reliable estimates for some of the procedures outlined in the manual. For example, the delay caused by "having one's analysis approved by a statistician" appears constant at 15 months (for both exploratory and confirmatory techniques). Further, "having a colleague (read: reviewer) read one's work" is worth on average 9 months (constant across the American and European continent). Contrary to expectation, "extending one's family" only buys 4.5 months (for baby boys). And even more surprisingly, maternity leave turns out to be ideal for finishing a book manuscript! The cause of this is, no doubt, my failure to select the right partner, i.e. one "with as little understanding for the nature of research work as possible". *Tack* Torkel *och* Knut!

Sheffield, July 2010

Contents

1. Introduction

Near-synonymy or near-sameness of meaning is an intruiging linguistic phenomenon. In a semiotic system characterized by limited lexical resources, near-sameness of meaning may seem aberrant: instead of increasing the expressive potential of the language by making a single lexical item express multiple meanings, near-synonymy decreases the language's expressive power by allowing several lexical items to convey (roughly) the same meaning. Example (1) illustrates how Russian features three different verbs for which dictionaries suggest 'try' as first English translation (and the situation is even more complex than this example would lead you to believe, as will become clear further on in this book).

(1)   Но Сирота все еще **силился** что-то сказать, и снова невозможно было понять ни слова из того, что он говорил. Малинин наконец не выдержал и прекратил эту обоюдную муку:
Ты **не старайся**, Сирота, все равно я не понимаю: у тебя рот разбитый ... Звук и только, а голоса нет. В госпитале полежишь – восстановится, а сейчас **не пробуй**, не мучь себя (. . .) [К. Симонов. Живые и мертвые.]

But Sirota was still **trying** to say something, and again it was impossible to understand a word of what he was saying. Finally, Malinin could not take it any longer and put an end to this mutual torture:
"**Don't you try**, Sirota, I can't understand you anyway: your mouth got smashed . . . . There is only sound, no voice. You'll be in hospital for a while – it will heal, but for now **don't try**, don't torture yourself" (. . .)[1]

Yet even if near-synonyms do name one and the same thing, they name it in different ways: they present different perspectives on a situation. Near-synonyms are neither in free variation, nor in complementary distribution. A clear understanding of the different import made by near-synonymous lexemes is thus relevant for the accurate and effective choice of words in communication.

Meaning being a rather elusive phenomenon, how can unsuspecting language learners detect and acquire the subtle differences between such words that are "very similar" in meaning? In this book, I am concerned with the clues language provides, thus leaving aside the role of other types of input, such as extralinguis-

---

1   All translations of text, originally written in Russian, are mine.

tic, socio-cultural input and contextual information gathered from experiencing situations in which the words are used. One of the linguistic mechanisms that can be used to make inferences about semantic facts, and one that plays a crucial part in this book, is syntax. The sentence structure surrounding a novel verb, and in particular the sets of syntactic formats a verb is used in, provides clues to that verb's meaning: it assists the learner in figuring out what the verb roughly means. In this sense, syntax acts as a kind of linguistic "zoom lens" (Fisher et al. 1991), helping the learner determine the verb's perspective on an event. The syntactic bootstrapping approach to language acquisition (Gleitman 1990: 27 and later publications) even claims that unobservable semantic properties must be marked in the syntax if the language learner is to discover what words in the language express them.

Yet the semantic information provided by sets of so-called subcategorization frames seems insufficient as a basis for acquiring *all* semantic generalizations (Gleitman 1990: 27, fn. 8). How would speakers distinguish between words that share most if not all frames such as, say, "break" verbs (Levin 1993, section 45.1), e.g., *break, chip, crack, crash, crush, fracture, rip, shatter, smash, snap, splinter, split, tear*? Any answer to this question has consequences for a theory of lexical organization. In this book, I pursue the hypothesis that the meaning of verbs in sentences might be parceled out (sometimes redundantly) between the clausal structure and the lexicon (cf. Fisher et al. 1991).

Specifically, I look into how language data of different degrees of abstractness, such as lexical items and grammatical constructions, interact in conveying and thus possibly assist in acquiring subtle meaning differences. In studying this division of labor between grammar and lexicon, I present an objective and verifiable, distribution-based solution to the three major problems of synonym research, i.e., delineation, internal structuring and description. More in particular, I concentrate on the following questions:

1. How can groups of near-synonyms be delineated? What kind of (intra-linguistic) information do speakers of a language have at their disposition to decide which words express similar meanings?
2. How are groups of near-synonyms structured internally? Which verbs are more similar to each other and how can similarity be measured?
3. How can highly similar lexemes be differentiated from each other? Is there a verifiable way to describe the scales of variation along which the synonyms differ? Is there an objective way to decide which properties discriminate best between semantically similar verbs?

The study of near-synonymy in itself provides interesting information about the structure of a fundamental lexical phenomenon, i.e., a sense-relation, as well as

about its semantic and related conceptual space. The account I present combines insights from different trends in cognitive linguistics and elaborates methods that allow measuring their theoretical claims against real data collected by means of elicitation tests with native speakers as well as extensive corpus research. The solutions I propose are based on in-depth case studies of near-synonymous verbs that express HAVING AN INTENTION, UNDERTAKING AN ATTEMPT and ACHIEVING RESULTS in Russian. These lexical items are ideal candidates for exploring the syntactic bootstrapping hypothesis, as the concepts they put into words are rather abstract, hence hardly observable. Furthermore, these verbs are likely to be acquired only later in life (cf. Dąbrowska 2009), making the discussion of whether syntactic bootstrapping applies from the earliest stages of language acquisition irrelevant here (for a brief yet comprehensive introduction to the issue, see Bowerman and Brown 2008).

But let us start with defining what near-synonyms are. In doing so, I contrast the dominant views in both the English (Cruse 1986, 2000) and Russian (Apresjan [1974] 1995; Apresjan et al. 1995) lexicological traditions.

1. What is near-synonymy?

The folk-definition of near-synonymy is deceptively simple: near-synonyms are different lexemes that express the same meaning or at least very similar meanings. Yet, what does "expressing the same meaning" really mean?

Traditionally, two words are considered synonymous in a sentence or linguistic context if the substitution of one for the other does not alter the truth value of the sentence. Two lexical units would be absolute synonyms if and only if all their contextual relations were identical. For this reason, it is commonly asserted that absolute, perfect or full synonyms do not exist. Synonyms, then, are defined as lexical items whose senses are identical in respect of "central" semantic traits, but differ in respect of so-called "minor" or "peripheral" traits.

Cruse (2000: 158–161) distinguishes two kinds of synonyms, i.e., "cognitive synonyms" and "plesionyms". If two lexical items are cognitive synonyms they must be identical in respect of propositional traits, but they may differ in respect of expressive traits. In other words, there can be differences of stylistic level and differences of presupposed field of discourse. Popular examples are *die, pass away* and *kick the bucket*. Plesionyms or near-synonyms, on the other hand, are more common. Plesionyms yield sentences with different truth-conditions: they are weakly contrastive, but the contrast does not destroy the synonymy. Permissible differences must, however, be either minor or backgrounded, and can be both. Minor distinctions are, for example, adjacent positions on scales

of degree, such as *fog* and *mist*, but also certain adverbial specializations of verbs, e.g., *amble* versus *stroll*, aspectual distinctions as between the state of being *calm* and the predisposition of being *placid* and differences of prototype center such as focus on physical factors for *brave* versus focus on intellectual and moral factors in *courageous*. Backgrounded major distinctions like gender can be found in the pair *pretty* and *handsome*. As the semantic distance between lexical items increases, plesionymy shades imperceptibly into non-synonymy.

Apresjan ([1974] 1995b: 223, 235ff) considers two words to be exact lexical synonyms or *točnye sinonimy* if they meet the following three requirements. Exact synonyms need to have totally coinciding interpretations[2], i.e., they translate into one and the same semantic expression. It is also required that exact synonyms show the same number of active semantic argument structure slots and that argument structure slots with the same number have identical roles. Finally, exact synonyms must belong to the same (deep) part of speech. Much more frequent than "exact" lexical synonyms are the *netočnye sinonimy*, non-exact or quasi-synonyms, in Cruse's terminology "near-synonyms". Quasi-synonyms differ from exact synonyms on the first criterion: quasi-synonyms are not characterized by totally, but by partially coinciding interpretations. Quasi-synonyms display two types of main differences. *Rodo-vidovye* differences arise when the meaning of one word is included in the meaning of the other, like the type of pain expressed by *sadnit'* 'smart, burn' that is included in *bolet'* 'ache, hurt'. *Vido-vidovye* differences are present when the meanings of two lexemes intersect, such as in *myt'* 'wash' vs. *stirat'* 'launder'.

How do we decide which words qualify as near-synonyms? Within the Western tradition, synonyms are defined contextually by means of diagnostic frames. As I mentioned before, popular examples of cognitive synonyms are *die, pass away* and *kick the bucket*; since these verbs only differ in expressive traits it is impossible to state **He kicked the bucket but he did not die*. Yet plesionyms differ in more than just expressive traits, so two plesionyms can be united in one sentence such as *He was killed, but I can assure you he was not murdered*.

In the Russian tradition, the decompositional approach prevails and synonyms are analyzed by means of a semantic metalanguage. Apresjan et al. (1995: 60, 2000: XL) defines the constitutive characteristic of "synonyms" as "the presence in their meaning of a sufficiently big overlapping part". To define the "sufficiency" of "big overlapping", the meanings of words are reformulated with the help of a special meta-language. The strict formulation prescriptions

---

2   It should be pointed out that Apresjan adheres to a "weak" interpretation of "total interchangeability" (Apresjan [1974] 1995: 219), which places his "exact synonyms" on the level of Cruse's "cognitive" or "propositional synonyms".

and the limited inventory of lexical primitives of this metalanguage facilitate comparison of meanings. The overlap has to be bigger than the sum of the differences for two lexemes, or at least equal to the sum of the differences in case of three or more lexemes. Apart from that, the overlap has to relate to the assertion of the definition that contains "genera proxima", the syntactic main word of which coincides.

2.   Theoretical framework: basic concepts

The research reported on in this book concentrates on near-synonyms from the perspective of formal onomasiological variation (coined as such by Geeraerts et al. 1994). In other words, it looks at alternative linguistic means used to designate the "same" concept or linguistic function.

Although near-synonymy is a fundamental phenomenon that influences the structure of our lexical knowledge, and has been recognized as such in Russian linguistics, it has received relatively little attention on this side of the Iron Curtain over the last few decades. Part of the reason for this might be that a graded, lexical phenomenon like near-synonymy does not fit in well with the theoretical frameworks that predominated Western linguistics during the second half of the 20th century. Cognitive linguistics is more suited than other frameworks to deal with a phenomenon characterized by high similarity and low contrastivity in meaning. In this book, I show that a combination of insights from different trends within cognitive linguistics is ideal for the study of near-synonymy. Furthermore, this theoretical apparatus enables me to provide an account of language data that is consistent with what is generally known about human cognition, an aim often referred to as "cognitive commitment" (Lakoff 1990).[3]

First of all, in cognitive linguistics and in Cognitive Grammar in particular, meaning comprises content as well as construal, i.e., the way in which content is construed (Section 2.1). Construal manifests itself at the level of choice of grammatical and lexical items alike. As a consequence, the difference is stressed in the semantic-cognitive import made by each construction as well as by each lexeme. Secondly, within cognitive linguistic and especially within

---

3   It has to be borne in mind that "cognitively plausible" does not necessarily imply "cognitively real". For some of the main claims made in this book I do present independent psycholinguistic validation. As concerns the non-validated results it needs to be stressed that care was taken to develop an objective and verifiable methodology. This should ensure that the results can be subjected to verification by researchers from neighboring disciplines, such as psycholinguistics.

Construction Grammar approaches, both constructions and lexemes, in fact, all linguistic units are considered form-meaning pairings that form larger sets or networks (Section 2.2). If both grammatical constructions and lexical elements are meaningful units, their meanings need to be compatible in order to yield felicitous combinations. Comparing the sets of constructions verbs can occur in then tells us something about the meaning components those verbs share. Thirdly, cognitive linguistics and Prototype Theory in particular promote explicating language structure in terms of other facets of cognition, i.e., they aim to identify cognitive capacities used to process entrenched form-meaning pairs. Particularly important for the graded phenomenon of near-synonymy is knowledge about human categorization mechanisms and the prototype structure of categories (Section 2.3).

2.1.   Cognitive Grammar and construal

As I stated before, in cognitively inspired approaches to language, grammar and lexicon are seen as forming a continuum and encoding meaning on different levels that are progressively characterized by a higher degree of specificity. The meaning conveyed by both types of structures is considered to be conceptual in nature, i.e., "experience is conceptualized in the process of encoding it and expressing it in language" (Croft 1999: 77 and 2007 for an elaborate discussion of the importance of the verbalization of experience for the emergence of grammar). This process of "construing the world" (Geeraerts 2006: 4), which results from general cognitive capacities, is known as construal. Human beings have the capacity to construe an experience or situation in alternate ways. Although differences in construal of an experience or situation may result from different perceptual, cognitive and expressive capacities as well as from the choices made by an individual speaker to profile certain aspects of an experience or situation in a given utterance, cognitive linguists stress the latter, viz. the difference in semantic-cognitive import made by the way in which content is presented:

> [T]he lexical and grammatical conventions of a language provide an array of alternative expressions for coding and conceptualization: expressions are often functionally equivalent but nonetheless different in meaning by virtue of the contrasting images they convey. (Langacker 1987a: 111)

Yet cognitive linguists are aware of the fact that construal is constrained by convention but also by the experience itself. Collective experience has given rise to a pool of structures from which individual speakers generally choose when formulating their experiences. Structure and experience are related: syntactic structure reflects semantic structure and the semantic structure corresponding to a syn-

tactic construction represents a conceptualization of experience (Croft 1999: 87–88, 2001: 128). In other words, all linguistic elements are form-meaning-experience triplets – hence the "experiential grounding of grammar" – that convey meaning ranging from highly schematic (i.e., grammatical units) to highly specific (i.e., lexical elements). And on a construction approach to language, form-meaning pairings on all levels and of any kind of abstractness and complexity are considered constructions.

2.2. Radical Construction Grammar and sets of constructions

Over the past two decades, constructions have made an impressive advance in linguistics. The conviction has been growing that the (grammatical) description of a language should consist of the analysis of its constructions, i.e., form and meaning pairings, the properties of which – both formal and functional – cannot be reduced to the component parts and their composition.

Croft's model (2001) is attractively non-reductionist. The "primitive construct" of Radical Construction Grammar is the construction, which is a complex entity. Constructions contain categories and relations, and these are defined by the constructions they appear in; they are not theoretical primitives. In other words, constructions are not derived from their parts, but instead the parts are derived from the constructions. Hence, the parts of a construction do not have an independent existence outside of the whole construction. Croft pleads for entirely abandoning the assumption that syntactic structures are made up of primitive categories and relations. He explicitly claims that the grammatical description of any language should exclusively consist of an analysis of the constructions in that language and the network of relationships among them. Likewise, a speaker's knowledge of a language is organized by the knowledge of relations between (the meanings of) constructions (van den Eynde 1995: 116ff; Goldberg 1995: 67; Croft 2001: 25–29). Constructions are thus non-derived patterns, abstract expressions in their own right, located at the schematic level of language; they feature open slots that can be filled up with lexicalized elements at a more specific level.

These constructions or grammatical patterns of a language are aptly described by means of the distributional method (Croft 1999: 69–74). On the distributional method, constructions define categories, albeit not in the strict sense. What matters is "the semantic interpretation of a word in a particular grammatical construction. By examining the meanings of verbs in constructions, we can establish semantic classes of events and conceptual meanings of constructions" (Croft 1998b: 91). This is precisely the route I take in this book.

2.3.   Prototype Theory and the structure of categories

The ability to create classes, i.e., the ability to categorize, is one of the fundamental qualities of human cognition that is most pervasively present in language.

> Categorization is not a matter to be taken lightly. There is nothing more basic than categorization to our thought, perception, action and speech (. . .) And any time we either produce or understand any utterance of any reasonable length, we are employing dozens if not hundreds of categories: categories of speech sounds, of words, of phrases and clauses, as well as conceptual categories. Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives. An understanding of how we categorize is central to any understanding of how we think and how we function, and therefore central to an understanding of what makes us human. (Lakoff 1987: 5–6)

Research on human categorization mechanisms is typically concerned with the formation and manipulation of non-linguistic psychological representations of a class of entities in the world, also known as concepts. Concepts fulfill an important function in linguistics as well, as they mediate the relation between a word and the range of entities in the world it may refer to. Conceptual approaches to meaning therefore identify (part of) the meaning of a word with the concept or concepts it gives access to in the cognitive system. And, as a consequence, "whatever is true of a concept must be true of the lexical item conveying that concept" (Murphy 2002: 391).

2.3.1.   *Horizontal and vertical category structure*

Concepts are considered to have the status of categories: they classify experience and give access to knowledge concerning entities, which fall into them. Three decades ago Eleanor Rosch refuted the psychological reality of an Aristotelian view on categories[4] by demonstrating the inadequacy of necessary and sufficient attributes for item classification. Instead, she presented a prototype approach to categorization, a probabilistic feature approach with instances displaying different degrees of representativity and similarity to a prototype. That prototype representation of a category is generally taken to be a generalization or abstraction of a class of instances falling into the same category (Hampton 1995: 104–105).

---

4   The Aristotelian use of the term "category" does not coincide, however, with the way "category" is used in Prototype Theory, nor with the interpretation it received later in cognitive linguistics (see e.g., Lakoff 1987); for Aristotle, *red* would belong to the category "quality" rather than being a category in itself (Ines Van Houtte, p.c.).

An important contribution to the informativity of prototypes for concepts seems to be made by the taxonomic level at which these concepts are situated (Ungerer and Schmid 1996: 97–98). In psychology (and consequently linguistics) three levels are recognized. Elements in categories on the superordinate level have no common Gestalt, one or very few category-wide attributes, but salient general attribute(s). Elements in categories on the basic level, on the contrary, do have a common Gestalt, and a large number of category-wide attributes. Subordinate categories, finally, display an almost identical Gestalt, a large number of category-wide attributes and salient specific attribute(s). In other words, categories become cognitively more economic and more informative on the basic level.[5] Hence, having prototypes on the basic level of a taxonomy ensures maximal information for minimal cognitive effort.

2.3.2. *Category structure representations*

What does the internal structure of a category that is organized by a prototype look like? Complex categories can be represented in two ways, i.e., as a schematic network or as a radial category.

According to Langacker (1987a: 369, 371), complex categories are best conceived and described as (hierarchical) schematic networks of interrelated senses. A schema is an abstract characterization that is fully compatible with all the members of the category it defines. These members of a linguistic category, e.g., interrelated senses, are linked to each other by categorizing relationships involving an act of comparison in which a standard is matched against a target.

Lakoff (1987: 83–84, ch. 6), on the other hand, promotes the (flat) radial category structure. A radial structure has, apart from a central instance, a number of conventionalized variations on that central instance. These variations are all understood as deviations from the central case. The non-central cases do not only have more properties than the central instance, they also have different properties. Moreover, non-central cases cannot be predicted by general rules and have to be learned since not all possible variations of the central case exist as subcategories: subcategories are culturally defined.

The two modes of categorization, i.e., schematic and radial categorization, are inherently related aspects of one and the same phenomenon and are often difficult to distinguish in practice (Langacker 1987a: 371 ff). There is a qualitative

---

5 Wierzbicka (1985: 243 note 1) points out that Rosch and her collaborators never explicitly stated whether any taxonomic level, subordinate, middle or superordinate, can function as basic level or whether this is always the "middle" level, but linguists have taken it to be the "middle" level in their writings.

difference, though. Schematicity-based categorization provides full sanction: the abstract characterization is fully compatible with all the members of the category it defines. Prototype-based categorization, on the other hand, provides only partial sanction: there are degrees of membership based on degrees of similarity, which fits near-synonymous concepts well.

3.   Methodology: a distribution-based approach

Although the larger cognitive linguistic framework offers the theoretical concepts needed to analyze near-synonyms, a methodology has to be developed for operationalizing those concepts and handling the data.

Traditionally, linguistic phenomena have been described in terms of (im)possibility, often using the minimal pair test. Unfortunately, a graded phenomenon like near-synonymy does not lend itself well to minimal pair tests; try and come up with a structure in which *try* is perfectly felicitous and *attempt* utterly unacceptable? In addition, analysts typically collect the data they need for this test by means of introspection. This approach has its drawbacks (for an overview see Gibbs 2006). To name but a few, when we take three different works (Černova 1996b; Apresjan et al. 1999; Evgen'eva 2001) that deal with verbs expressing INTENTION, ATTEMPT or RESULT in Russian, we see that these dictionaries do not list the same verbs, they do not group the verbs listed in the same way and they do not necessarily interpret the verbs in the same way. These problems can be overcome, as I illustrate in this book. Taken together, the case studies I present provide an objective and verifiable methodology that can be applied to delineating, structuring and describing groups of near-synonyms in Russian as well as in other languages.

A thorough analysis of elicited data on the set or network of constructions a verb takes part in provides a verifiable solution to the problem of *delineating* groups of near-synonyms (cf. Apresjan 1967 for Russian; Eggermont and Melis 1992 for French; Levin 1993 for English; Schøsler and Van Durme 1996 for Danish). On a Construction Grammar approach to language both constructions and lexemes have meaning; as a consequence, the lexeme's meaning has to be compatible with the meaning of the construction in which it occurs and of the constructional slot it occupies to yield a felicitous combination. Therefore, the range of constructions a given verb is used in and the meaning of each of those constructions are revealing of the coarse-grained meaning contours of that verb. The results can then be used to delineate groups of near-synonymous verbs.

On this approach, near-synonyms share constructional properties, even though the extent to which a construction is typical for a given verb may vary

and the individual lexemes differ as to how they are used within the shared constructional frames. A fine-grained corpus-based analysis solves the problem of *structuring* a group of near-synonymous lexemes internally, as well as of *describing* the prototype center for each lexeme. Working on the basis of a large collection of texts that is representative, at least for a particular type of language use (and even here we are dealing with a collection of idiolects from which we assume conventionalized generalizations are made), gives an idea of what is normal and what deviates from the norm. Also, corpora facilitate observing several variables simultaneously and judging them in terms of probability instead of (im)possibility, which seems beneficial in the case of a graded phenomenon like near-synonymy.

In order to approach meaning from a probabilistic point of view, i.e., compare verbs in an objective and verifiable way, their meaning needs to be quantified. In doing so, I make use of the distributional hypothesis, assuming similarity between meaning and (frequency) distribution: words that mean similar things are used in a similar way. To describe that usage, I annotate a representative sample of corpus extractions with a system of labels that is referred to as "ID tags" and yields a "Behavioral Profile" of the structural and lexical preferences each verb has. Performing statistical analyses on these Behavioral Profiles makes it possible to reveal the category structure and visualize the position of and relative distance between the verbs in a category. Furthermore, the results are used to provide an explicit representation of how each node prototypically specifies a view on a situation and in which respects the near-synonyms differ. Throughout this book, I focus on the interpretation of the statistical results. For linguists, statistics is a means, not an end in itself. It is merely a tool, albeit a handy one, that makes exploring large datasets easier and facilitates extrapolating findings to the larger population.

The proposed usage-based approach is compatible with one of the basic tenets of cognitive linguistics. When native speakers of a language say that they know a word, they typically mean that they know how to use it in everyday discourse, not that they are able to cite its dictionary definition. People learn how to use words by observing how words are used together in phrases and sentences, i.e. contexts. Even stronger, people can learn how words are used on the basis of information provided by purely linguistic contexts (Miller and Charles 1991: 4). If meanings are contextual relations, the similarity of meaning and the similarity of contexts must covary (Miller and Charles 1991: 9). In other words, the difference in semantics between, say, two types of verbs emerges in the speaker's grammar as a result of differences in usage, e.g., some types of [V_{FIN} V_{INF}] constructions are never rephrased as [V_{FIN} Pronoun], as we will see in the next chapter. One caveat is in order regarding context: I limit my

attention to those elements found within sentence boundaries. Socio-linguistic variables are not taken into account, yet the methodology I propose can easily accommodate these factors provided that sufficiently large amounts of data on the variables of interest are available (see Glynn 2010).

The methodology I present thus operationalizes insights from cognitive linguistics and measures them up against real data, while presenting a viable alternative to introspection-based approaches to linguistic categorization and lexical description. The introspection-based approaches lack a precise measure to determine both the degree of relatedness among near-synonyms and the importance of each of the patterns and variables encountered for the description of the differences between near-synonyms. At the same time, the Behavioral Profile approach I propose to use when preparing the non-elicited linguistic data for quantitative analysis is innovative for the field of corpus linguistics, with which it has methodological affinities, and it can be seen as a step towards the operationalization of the elusive concept meaning. On a theoretical level, the analyses I present provide insight into the workings of two of the most fundamental facets of language, grammar and lexicon. A radical distributional approach facilitates revealing how powerful the form-meaning relationship really is and how semantic knowledge is parceled out between grammar and lexicon.

4.   Outline of this book

In Part One of this book (Chapters Two and Three) I present a construction-based method for delineating semantically coherent groups of verbs, working from the assumption that the closer any two verbs are in their semantic structure, the greater the overlap should be in their licensed syntactic structures. I bring that method to bear on the differences between six lexical items a native speaker of Russian has at his/her disposal to express 'having an intention' and 'having a plan'; the results I present are based on data from elicitation tests with 15 native speakers of Russian.

The methodology I propose for delineating groups of near-synonyms builds on the fact that human beings have the capacity to construe a situation in alternate ways; hence, knowledge of one construction involves simultaneous knowledge of a complex network of constructions. Given that each such construal or construction has meaning, albeit rather schematic meaning, with which the meaning of the lexemes used has to be compatible, I endorse the view that a set or network of constructions reveals the coarse-grained meaning contours of the verb that appears in it. The alternating constructions I use (Chapter Two) highlight three different facets of the event described as they are encoded in the argu-

ment structure (does the infinitive belong to the argument structure of the finite verb? are there *that*-complementation alternatives?) and event-temporal properties (as expressed in aspectual preferences and restrictions on combinability with temporal adverbs) of the finite verbs. The differences between the networks of constructions characterizing each verb play an important role in distinguishing between verbs (Chapter Three) that express plans (*planirovat'* 'plan') and intentions (*predpolagat'* 'intend, propose', *dumat'* 'intend, think (of)', *namerevat'sja* 'intend, mean', *sobirat'sja* 'intend, be about' and *xotet'* 'want, intend').

In Part Two of this book (Chapters Four and Five) I present an objective and verifiable corpus-based methodology that incorporates and operationalizes findings from Prototype Semantics for analyzing nine near-synonymous verbs that express ATTEMPT in Russian, i.e., *probovat'* 'try', *pytat'sja* 'try, attempt', *starat'sja* 'try, endeavor', *silit'sja* 'try, make efforts', *norovit'* 'try, strive to, aim at', *poryvat'sja* 'try, endeavor', *tščit'sja* 'try, endeavor', *pyžit'sja* 'try, go all out' and *tužit'sja* 'try, make an effort, exert oneself'.

Exploratory statistical analysis of an enormous amount of manually coded corpus data (in total 137,895 data points), provides a verifiable basis for visualizing the radial structure of the category that includes the nine verbs: it identifies the central cluster as well as peripheral clusters. In addition, it facilitates elucidating the elaborative distance between the verbs in the network: the greater the distance, the less synonymous the verbs are (Chapter Four). Finally, descriptive statistics computed from the corpus-based Behavioral Profiles reveal in which respects two or more verbs resemble each other or differ from each other. Frequency information from Behavioral Profiles tells us precisely how the near-synonyms convey a different view on a particular situation or an "alternate 'window' on a common knowledge base" (Langacker 1987a: 378) and in which respects the peripheral nodes supplement the meaning of the central nodes. In other words, these results give quantitative depth to and provide an objective basis for the qualitative analysis or "lexical portraying" (Apresjan et al. 1995) of the prototype center of each of the verbs (Chapter Five), and yield insight into how the differences in prototype center between the central nodes are grounded in human experience of reality.

In Part Three (Chapter Six), I evaluate the distributional approach to near-synonymy as well as the suggested methods from four different perspectives. First, I focus on establishing how well the Behavioral Profile approach performs, i.e., how well the selected variables are able to predict and account for the studied phenomenon. Next, I show that discovering the internal structure of a group of near-synonyms aids in understanding why certain constructional alternatives are available for some verbs and not for others. Then, I look into the question of which differences, abstract grammatical or concrete lexical ones, approximate

best the intuitions native speakers have. And finally, I compare the results of the presented analysis of non-elicited data with elicited data that reveal native speaker intuitions on the degree of closeness between the verbs in each group.

By means of the case-studies presented in this book, I aim to provide at least partial answers to common questions like "Where do near-synonyms come from and why do we have them?" or "Which are "good" near-synonyms and is there reason to assume that they come in pairs?" Answers are formulated on the basis of data collected and analyzed using a methodology that identifies patterns of usage at different levels of granularity, thus progressively refining the type of meaning taken into consideration. This approach has the potential to form the basis for an objective semantic analysis, the ultimate test-case being that of near-synonymy, tackled in this book.

In sum, this monograph makes scientific contributions at three different levels, the descriptive, methodological and theoretical. It provides a detailed account of the category of complex events as expressed by $[V_{FIN} \, V_{INF}]$ structures in Russian: starting from the coarse-grained constructional level and gradually working its way down to the fine-grained lexical level it reveals how complex events are structured in Russian and how the 300 odd verbs that participate in the $[V_{FIN} \, V_{INF}]$ structure relate to each other. On the basis of several detailed case-studies on near-synonymous complex events it re-appraises traditional semantic theory and re-evaluates existing accounts of near-synonymy: it demonstrates that near-synonymy is adequately dealt with in a cognitive linguistic framework, that cognitively-oriented analyses of near-synonymy benefit from taking a corpus-based, i.e., usage-based, perspective, and that grammar and lexicon are two extremes of one and the same form-meaning continuum that "conspire" (Wierzbicka 1988) to convey meaning. These findings shed light on the overarching question of what language in use has to offer the learner in his/her quest for the meaning of lexical items and grammatical constructions.

2. Degrees of event integration: delineating groups of near-synonyms

1. Are all [V$_{FIN}$ V$_{INF}$] structures created equal?

In this chapter I present an objective and repeatable, construction-based method for delineating semantically coherent groups of verbs. In doing so I work from the assumption that the closer any two verbs are in their semantic structure, the greater the overlap should be in their licensed syntactic structures. As there is a limit to the alternate grammatical coding options available per situation and thus per verb, data on constructional restrictions and preferences can be used to delineate groups of near-synonymous verbs.

Within Cognitive Grammar, basic grammatical classes such as noun and verb are considered as symbolic structures (Langacker 1987a: 189). As they share fundamental semantic properties, their semantic poles instantiate a single abstract schema that can be characterized. When two or more symbolic structures integrate, viz., combine to form a more elaborate expression on the syntagmatic plane, a grammatical valence relation comes into existence. The valence properties of the building blocks can only be explained and understood with reference to their internal structure (Langacker 1987a: 277). Important here is to bear in mind that in that process of integration, one component may need to be adjusted in certain details when integrated with another to form a composite structure, thus resulting in accommodation (Langacker 1987a: 75–76). This, I argue, may in some (extreme) cases lead to re-analysis of the construction as a whole.

Integration, accommodation and the internal structural differences between seemingly identical external structures are at the center of attention in this chapter, in particular in Section 2.1. A fine illustration of how constructions outline the coarse-grained meaning contours of the verbs that occur in them is provided by the six lexical items a native speaker of Russian has at his/her disposal to express his/her plans (*planirovat'* 'plan' and intentions (*predpolagat'* 'intend, propose', *dumat'* 'intend, think (of)', *namerevat'sja* 'intend, mean', *sobirat'sja* 'intend, be about' and *xotet'* 'want, intend'). This case study is presented in detail in Chapter 3.

1.1.   Near-synonyms: between grammar and lexicon?

In Russian, verbs that express plans and intentions have in common that they combine with an infinitive, for example *I _intend_ _to write_ a book about near-synonyms*. In all, there are 293 verbs that occur in this pattern, henceforth [V_{FIN} V_{INF}]. This construction is the point at which all verbs analyzed in this book intersect, and all concepts that can be expressed using the [V_{FIN} V_{INF}] are included. Given that within cognitive linguistics abstract grammatical structures are assumed to have meaning too, this communality means something. On a motivational iconic view, the [V_{FIN} V_{INF}] pattern signals a high degree of integration between the events expressed; hence, all verbs that appear in this structure as finite verbs share at least the meaning component that facilitates integration with a second event. Yet, there seem to be differences in the degree to which the events are integrated and in this chapter I propose three parameters to elucidate these degrees of integration (cf. Divjak 2007a).

All three parameters relate to the function verbs typically fulfill. Verbs express events or situations that have participants and take place at certain moments in time. The main participants of events or situations are encoded in the verb's argument structure that is the basis of the simplex sentence. Therefore, first of all, I investigate whether the infinitive event merely participates in the finite verb event and can be accommodated under one of the argument structure slots the finite verb opens up. Next, I consider the fact that, if two (or more) events are being reported on, the most usual way to link up the two verbs expressing these events is by means of two coordinated main clauses or complex sentences consisting of a main clause and a complement or subordinate clause. Here I focus on the *that*-clause alternatives that are available for rephrasing the infinitive event. Finally, the events both verbs express take place at a specific moment in time; yet in some cases the events expressed in [V_{FIN} V_{INF}] structures can take place at different moments in time whereas in other cases co-temporality or tight sequentiality is required.[6] In addition, events have a specific temporal contour, i.e., an imperfective and/or perfective aspect, and I investigate how aspectual properties interact with temporal separability.

Each of these three parameters, i.e., argument structure, *that*-complementation and temporal separability, has cognitive-semantic dimensions, i.e., reification (Langacker 1987a), objectification (Wierzbicka 1988; Langacker 1991)

---

6   Although adverbial specifications of location can be expected to display similar behavior, they are not focused on: different from temporal information, locational information is not expressed on the verb in Russian and can therefore be considered secondary to temporal information in expressing event structure, at least for verbs that do not express motion.

and distancing (Givón 2001). Taken together, these three dimensions can be interpreted as encoding the degree of integration between the finite verb event and the infinitive event. Given that the results of the analysis show that there is evidence for a cline with eight different marks of integration between the events expressed by means of a [V_{FIN} V_{INF}] pattern (cf. Givón's "binding scale"), interpretation of the findings suggests that these eight marks on the scale illustrate the degree of independence both verbs and the events they express have with respect to each other.

1.2. Theoretical motivations for taking a construction-based approach to delineation

In the literature two ways of delineating near-synonyms dominate the scene. On the one hand, there are the traditional alternatives for delineating near-synonyms; in order to "measure" similarity, these approaches rely on intuition (frequently applied in dictionaries, e.g., Evgen'eva 2001) or on rephrasing lexemes into a metalanguage made up of semantic primitives (Apresjan et al. 1995 and subsequent publications). Natural Language Processing applications, on the other hand, favor the automatic extraction of near-synonyms based on information included in machine-readable dictionaries of, say, the WordNet type, or on a comparison of distributional patterns extracted from corpus data. These patterns include reference to (a certain number of) words surrounding the target word(s), and/or grammatical dependency relations typical of the target word(s). The latter approach, that is in line with what I propose in this chapter, seems to yield particulary good results. Peirsman et al. (2007: 15–16) found that a full syntactic content model outperforms models based on co-occurrences, thus strengthening Miller and Charles' (1991) earlier findings. Miller and Charles (1991: 11) reported that measures of contextual similarity based on co-occurrence confirmed the contextual hypothesis only for short distances in semantic space. Furthermore, co-occurrence measures dismember the contexts they are supposed to represent, and therefore do not approach this task the way human beings do since a word's contextual representation is more than just a collection of words (Miller and Charles 1991: 23).

In this book, I focus on one particular type of distributional information, viz., information contained in sets of argument structure constructions. It is now commonplace in cognitive linguistic literature to assume that the constructions a specific lexeme or lemma can occur in correlate with the semantic characteristics of that lexeme or lemma. Yet the idea itself has a long history. In 1967 Apresjan published an "experimental investigation into the semantics of Russian verbs", to name but one example relevant for Russian. Syntactic "transforms" were