UiT The Arctic University of Norway

Faculty of Humanities, Social Sciences and Education

Adjectivization in Russian

Analyzing participles by means of lexical frequency and constraint grammar

Uliana Petrunina

A dissertation for the degree of philosophiae doctor – February 2021

Adjectivization in Russian

Analyzing participles by means of lexical frequency and constraint grammar

Uliana Petrunina

A dissertation submitted for the degree of
philosophiae doctor (PhD)

Faculty of Humanities, Social Sciences and Education
UiT: The Arctic University of Norway
Norway
February 2, 2021

Contents

Abstract xv

Acknowledgments xvii

Abbreviations xix

Note on Transliteration and Glosses xxi

1 Introduction 1

1 Motivation . . . 1

2 Scope . . . 2

3 Background . . . 3

4 Relevance . . . 4

5 Outline . . . 4

2 Methodology overview 7

1 Introduction . . . 7

2 Weighted finite-state transducer . . . 8

2.1 Basic definitions . . . 8

2.2 Applications of weights and weighted FSTs . . . 10

3 Morphological annotation and disambiguation . . . 11

3.1 Morphological transducer . . . 12

3.2 Weights in morphological transducers . . . 13

3.3 Lexicon compiler . . . 14

3.4 Morphological analyzer and lexicons . . . 16

4 Constraint Grammar . . . 19

4.1 Background to CG formalism . . . 20

4.2 Russian CG . . . 21

4.2.1 Lists and sets . . . 22

4.2.2 Constraints . . . 23

4.2.3 Basic examples of rules from the Russian CG . . . 24

5 Summary 25

3 Ambiguity and adjectivization 27

1 Introduction 27

2 Approaches to ambiguity 28

2.1 Overview of ambiguity 28

2.2 Estimation and classification of ambiguity 30

3 Conversion 32

3.1 Conversion against the background of word formation in Russian 33

3.2 Overview of conversion 34

3.2.1 Properties and types of conversion 34

3.3 More about morphosyntactic properties of conversion 36

3.3.1 Conversion versus zero derivation 37

3.4 Summary 38

4 Approaches to adjectivization 40

4.1 Syntactic approach 41

4.1.1 Adjectivization based in syntax 42

4.1.2 Marginal cases of adjectivization 47

4.1.3 Predicativity and attributivity 49

4.1.4 Classification of participles 53

4.1.5 Summary 55

4.2 Morphosemantic approach 56

4.2.1 Adjectivization based in semantics 57

4.2.2 Causes and stages of adjectivization 61

4.3 Summary 70

5 Conclusion 71

4 Exploratory analysis 75

1 Introduction 75

2 The adverbial *očen′* ‘very’ construction 77

2.1 Data description 78

2.2 Data analysis and interpretation 79

2.2.1 Overview of the distributions 79

2.2.2 Overview of the semantic properties of the lemmas in the most frequent constructions 81

2.2.3 Statistical analysis of the ratio distributions and semantic classes 83

2.2.4 Context analysis of the lemmas in the most frequent constructions 89

2.3 Discussion and summary 95

3 Morphosyntactic properties and corpus frequency . . . 96
3.1 Data and sources . . . 98
3.2 Pervasiveness of participles . . . 99
3.2.1 High-frequency range . . . 100
3.2.2 Mid-frequency range . . . 105
3.2.3 Low-frequency range . . . 106
3.2.4 Summary . . . 107
3.3 Statistical analysis of the distributions . . . 108
3.3.1 Experiment design: Data and definitions of models . . . 109
3.3.2 Overview of the distributions . . . 111
3.3.3 Statistical modeling . . . 116
3.3.4 *GLM* model performance . . . 119
3.3.5 Test of the fit . . . 121
3.4 Summary . . . 123
4 Conclusion . . . 124

5 Disambiguation 127
1 Introduction . . . 127
2 Weighting . . . 129
2.1 Background to weighting . . . 129
2.1.1 Implementation of weights . . . 129
2.1.2 Transformation of weights . . . 130
2.2 Weighting the Russian finite-state transducer . . . 131
2.2.1 Data description . . . 132
2.2.2 Transformation of weights . . . 133
2.2.3 Implementation of weights . . . 135
2.2.4 Summary . . . 137
3 Development of the gold standard . . . 137
3.0.1 Data description . . . 138
3.0.2 Selection of sentences for the gold standard . . . 138
3.0.3 Preparation of examples for survey . . . 139
3.0.4 Survey design . . . 141
3.0.5 Interpretation of survey results . . . 142
3.0.6 Summary . . . 144
4 Disambiguation experiment . . . 145
4.1 Disambiguation models . . . 145
4.2 Design and description of CG rules . . . 148
4.2.1 Overview of CG rules . . . 148
4.2.2 Test corpora . . . 152

4.2.3 Unweighted CG rules 153
4.2.4 Weighted CG rules 154
4.2.5 Factors of adjectivization and weights in CG rules 156
4.3 Summary 162
5 Evaluation 162
5.1 Evaluation metrics 163
5.2 Unresolved ambiguity 165
5.2.1 Use of the CG weighted rule in the MAX model 168
5.3 Evaluation methods 170
5.3.1 Evaluation 1 171
5.3.2 Evaluation 2 173
5.4 Model performance 174
5.5 Summary 177
6 Conclusion 178

6 Conclusions **181**
1 Summary 182
2 Implications 186
3 Future directions 189

Bibliography **193**

Appendices **206**

A Methodology **207**
1 Basic definitions for weighted transducers 207
2 Files and commands 208

B Ambiguity **209**

C Conversion **211**

D Adverbial *očen′* construction **212**
1 Distributions 212
2 Semantic classes 214
3 *Očen′* constructions with the highest ratio 216

E Statistical analysis: adverbial *očen′* construction **217**

F Distribution analysis **220**

G Statistical analysis: frequency distributions **223**

1 Binary logistic regression models 223

1.1 Model 1: frequency, ratio 223

1.2 Model 2: tense, voice, frequency 224

1.3 Model 3: aspect, transitivity, frequency 226

H Survey design **228**

1 Survey report 228

2 Ambiguous sentences 235

I CG rules for ambiguous participles **236**

List of Tables

1 Correspondence table for Cyrillic and transliterated letters. xxi

2.1 An example of a cohort that contains the ambiguous base forms *на* ‘on/there’ used as an interjection and as a preposition. 13
2.2 Generated morphological analysis for the base form *слово* /slovo/ ‘word’. . . . 13
2.3 The weighted morphological analysis of the sentence ‘*a student reads books*’. . 16
2.4 Elements in the first line of the morphological analysis and their definitions. . . 19
2.5 The elements specifying the CG rule. 23
2.6 The elements specifying CONTEXT. 24

3.1 Summary of the properties of conversion. 34
3.2 Classification of participles with regard to their semantic and morphosyntactic properties that favor or disfavor adjectivization, based on Kalakuckaja’s (1971) and Kustova’s (2012) classifications, with some additional information from Černega (2009). 64
3.3 Classification of semantic and morphosyntactic properties favoring and disfavoring adjectivization. 68

4.1 Factors of adjectivization investigated in the exploratory analysis. 75
4.2 Overview of counts in the Araneum corpus. 78
4.3 The lexicon size of the Araneum corpus. 78
4.4 A sample of 22 *očen′* + PRESP and *očen′* + PRS V constructions and their ratio values. The largest ratio value in each row is in bold. 81
4.5 Semantic classes used for the more frequent verbal constructions and participial constructions. 82
4.6 Counts of semantic classes. 84
4.7 A sample of the eight *očen′* + PRESP and *očen′* + PRS V constructions with the highest frequency among the other constructions in the entire ratio distribution. The largest ratio value in each row is in bold, and the decimals are rounded to four places. 90

4.8 Types of verbs and the number of the morphological types of the participles they form. Morphological types refer to the combination of transitivity and aspect features a participle can represent, for example, a transitive imperfective participle. 100

4.9 Groups of lemmas in the mid-frequency interval: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V). 101

4.10 The top ten most frequent verbal lemmas. 102

4.11 Groups of lemmas in the mid-frequency range expressed in percentages: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V). 105

4.12 Groups of lemmas in the low-frequency range: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V). 107

4.13 Size of the datasets, expressed in number of rows (N). Each dataset contains variables used in the statistical experiment. For example, tense and voice variables are grouped as *tense.voice* in Dataset 2. 109

4.14 Estimate of variables in the datasets 110

4.15 Description of the variables. 111

4.16 Estimated regression parameters, standard errors, and p-values for the binary *GLM* Models 1, 2, and 3. 120

4.17 *KL* R^2 scores for Models 1, 2 and 3. 121

4.18 *LRT anova* test for *GLM* Model 1, 2 and 3. 122

5.1 The Zipf scale of word frequency (van Heuven et al., 2014: 1180). 131

5.2 Estimates for the I-RU lexicon for components of the corpus. 133

5.3 The structure of the frequency list including the frequency rank, raw frequencies, the lemmas *настоящий* ‘real’, *фабрика* ‘factory’, *расходиться* ‘part’, *подтвердить* ‘confirm’, and their tags ordered by the frequency rank of the lemmas. 133

5.4 Types of value transformation for the lemmas that correspond to participial word forms in the verbal lexicon. 134

5.5 The structure of morphological readings with weights assigned by the morphological analyzer. The structure consists of a word form, a morphological analysis of each POS, and the weight associated with it. 136

5.6 Estimates for the components of the SynTagRus dependency treebank. 138

5.7 Counts of word forms and sentences in two subparts of the gold standard, primary annotation and survey annotation. 139

5.9 Overview of ambiguous and control examples. 140

5.10 The number of participial, adjectival and remaining ambiguous word forms after the respondents' annotation. 144
5.11 Disambiguation of CG models defined on the basis of syntactic context and weights. $\mathrm{W_M}$ and $\mathrm{W_{MC}}$ are components added to the CTX and MAX models. GSD is a baseline model trained on the Russian UD treebank using UDPipe. . 146
5.12 Operations and notation of rules for specific contexts. Weighted rules are noted as *WPTCP* while unweighted rules, as *PTCP*. 150
5.13 Weighted and unweighted rules describing syntactic and morphological factors. Numbers are given in raw counts and percentages. 151
5.14 Factors of adjectivization defined in the CG rules used for disambiguating participles, both in specific and general contexts. 152
5.15 Syntactic characteristics for unweighted CG rules that describe specific context. 158
5.16 Syntactic characteristics for unweighted and weighted CG rules that describe general context. 159
5.17 Set of weighted rules and the morphosyntactic properties that they describe. . . 161
5.18 Confusion matrix with classification metrics. 163
5.19 Summary report for the metric scores in the disambiguation experiment for the CTX model, Evaluation 1. 172
5.20 Summary report for the metric scores for the CTX model in the disambiguation experiment, Evaluation 2. 174
5.21 Performance of the models for classifying adjectives (ADJ) and participles (PTCP), Evaluation 1. Metrics are rounded to two decimal places. 174
5.22 Overall performance of the models, Evaluation 1. Metrics are rounded to two decimal places. 175
5.23 Overall performance of the models, Evaluation 2. Metrics are rounded to two decimal places. 176

A.1 Types of semirings and interpretations of the weight of the global path, as well as its binary addition ⊕ and multiplication ⊗ operations, which can be used as other types of operations depending on the type of semiring and the interpretation of the global path weight. The weight of transition represents the value of the weight selected by the algorithm using a given semiring. 207

B.1 Classification of morphosyntactic ambiguity. 210

C.1 Manova's (2011) classification of conversion lists the linguistic level, the types of conversion, and properties of these types illustrated by examples. 211

D.1 The complete distribution of participial/verbal word forms and the *očen′* constructions with which they combine. 214

D.2 The list of the lemmas annotated with the tags indicating semantic classes. . . . 216
D.3 A sample of the eight constructions of *očen′* + PRESP and *očen′* + PRS V that had the highest frequency among other constructions in the list presented in Table D.1. 216

F.1 Suffixes for the full forms of participles; that is, present/past active/passive forms.220
F.2 Suffixes for the short forms of participles; that is, present/past active/passive forms.220
F.3 Forms of inflection in the declension of passive and active participles. 220
F.4 Verbal lemmas with no corresponding participles within the range of 1–100. . . 222
F.5 The list of verbal lemmas without corresponding participles 222
F.6 List of verbal lemmas without corresponding participles 222

H.1 Classified sentences with the tags *ptcp*, *adj*, and *ambig* assigned to the ambiguous word forms after the completion of the survey. 235

I.1 Rules for resolving the ambiguity of participles available in the Russian CG. The rules are presented in the order in which they appear in the Russian CG. 246

List of Figures

2.1 An example of an undirected graph G with the vertices a, b, c, d, e. 8

2.2 An example of a labeled directed graph over the alphabet A. The labels are the letters a and b. 9

2.3 A sequential transducer on the input $abaa$ and the output 01001 (Pin, 2016). The symbol ε represents an empty string. 9

2.4 An example of a WFST. 10

2.5 Morphological analysis of the sentence. 12

2.6 An example of a regular expression that produces a transducer that maps the strings *adb* onto *acd* with the weight of 1.0. 13

2.7 A simple weighted English lexicon. 15

2.8 An example of entries from the verbal lexicon for the Russian morphological analyzer. 16

2.9 An example of entries from the adjectival lexicon for the Russian morphological analyzer. 17

2.10 An example of entries from the verbal lexicon for the Russian morphological analyzer. 17

2.11 An example of a cohort for the ambiguous participial word form *следующий* /sledujuščij/ ‘next-ADJ, following-PTCP’ provided by the Russian morphological analyzer. Information about the POS tags *+V* (verbal forms) and *+A* (adjectival forms) is specified in the verbal and adjectival lexicons of the analyzer. The tag *PrsAct* corresponds to present active participles. 18

2.12 An example of LIST entries from the Russian CG. 22

2.13 An example of SET entries from the Russian CG. 22

2.14 Elements that constitute a simple CG rule. 23

3.1 Scheme of adjectivization. 69

4.1 Distribution of lemmas and *očen′* constructions in the Araneum corpus. 80

4.2 The percentages of the semantic classes for the interval with more frequent *očen′* + PRS V constructions (1–56). 82

4.3 The percentages of the semantic classes for the interval with more frequent *očen′* + PRESP constructions (57–90). 83

4.4 The semantic classes of the base verbs across the ratio distributions of the constructions *očen′* + PRESP. 85

4.5 The semantic classes of the base verbs across the ratio distributions of the constructions *očen′* + PRS V. 86

4.6 Verbal and participial lemmas ordered by the frequency rank of verbal lemmas (interval 1–2160), and the ipm frequency of verbal lemmas. 101

4.7 Verbal and participial lemmas ordered according to the rank of the verbal lemmas' frequency (interval 2161–4320). 105

4.8 Verbal and participial lemmas ordered according to the rank of the verbal lemmas' frequency (interval 4321–6480). 106

4.9 The distribution of ambiguous and unambiguous participial lemmas across the ratio of participial lemmas to verbal lemmas, and the ipm frequency of verbal lemmas. 112

4.10 Ipm frequency interval of 4–4.7. 113

4.11 Violin plots displaying the distribution of the ratios of participles grouped according to ambiguity (0 and 1), tense and voice, and ordered by their ratio to finite/infinitive verbs. 114

4.12 Violin plots displaying the distribution of the ratios of participles grouped according to ambiguity (0 and 1), transitivity and aspect, and ordered by their ratio to finite/infinitive verbs. 115

4.13 Binominal distribution in GLM. 118

5.1 A sample of the list consisting of participial lemmas (1st column) and their corresponding weights (2nd column) obtained after the Zipf transformation of the corpus frequencies. 135

5.2 Distribution of weighted and unweighted lemmas across the lexicons *verbs.lexc* and *adjectives.lexc*. 136

5.3 Counts of morphological types of participles in the ambiguous and control examples. 141

5.4 The counts of tags (participle as *ptcp*, adjective as *adj* and ambiguous as *ambig*) assigned to the respondents' responses for the ambiguous examples. 143

5.5 An example of the sentence and a word form annotated by the GSD model. The underlined, unambiguous word form *населенных* /naselennyx/'service', as in *населенных пунктов* 'residential areas, settlements' is annotated as a participle with the tags *ADJ JJL*. 147

5.6 Percent distribution of unweighted and weighted rules written for disambiguating participles in the Russian CG. 149

5.7 Percent distribution of rules with specific and general contexts, and also rules that do not describe any context around an ambiguous word form. 150

5.8 The cohort of morphological readings for the output of the word form *расширенного* ‘extended’ after disambiguation using the rule *REMOVE:WPTCP-A7*. 156

5.9 The cohort of morphological readings for the word form *населенных* ‘residential’ output after disambiguation using the rule *SELECT:maxweight*. The adjectival readings selected by the rule are underlined. 156

5.10 Ambiguity rate across the disambiguation models. 166

5.11 Percentage ratio of ambiguous word forms across the disambiguation models. . 166

5.12 Output of the CTX model wherein the participial and adjectival readings were not disambiguated. 167

5.13 Output of the MAX model wherein the participial and adjectival readings were not disambiguated. 167

5.14 Output of the MAX model with missing adjectival readings in the cohort. . . . 168

5.15 The cohort of the morphological readings with weights for the word-from *открытый* ‘open’ before disambiguation. 169

5.16 The cohort of the morphological readings with weights for the word form *открытый* ‘open’ before disambiguation. 169

5.17 The cohort of the morphological readings with weights for the word-from *открытый* ‘open’ before disambiguation. 169

5.18 The cohort of the morphological readings with weights for the word form *открытый* ‘open’ before disambiguation. 170

5.19 Confusion matrix in Evaluation 1, classification of adjectives. 171

5.20 The scheme for a confusion matrix for the classification of participles in Evaluation 1. The corresponding confusion matrix is given in Figure 5.19b. 172

5.21 Confusion matrix for the CTX model in Evaluation 2. 173

F.1 Distribution of verbal and participial lemmas depending on the frequency ranks of verbal lemmas. 221

F.2 Ratio of imperfective and perfective participial lemmas to verbal lemmas distributed across the ranks of the verbal lemmas. 221

Abstract

This dissertation explores the factors that restrict and facilitate adjectivization in Russian, an affixless part-of-speech change leading to ambiguity between participles and adjectives. I develop a theoretical framework based on major approaches to adjectivization, and assess the effect of the factors on ambiguity in the empirical data. I build a linguistic model using the Constraint Grammar formalism. The model utilizes the factors of adjectivization and corpus frequencies as formal constraints for differentiating between participles and adjectives in a disambiguation task.

The main question that is explored in this dissertation is which linguistic factors allow for the differentiation between adjectivized and unambiguous participles. Another question concerns which factors, syntactic or morphological, predict ambiguity in the corpus data and resolve it in the disambiguation model. In the theoretical framework, the syntactic context signals whether a participle is adjectivized, whereas internal morphosemantic properties (that is, tense, voice, and lexical meaning) cause or prevent adjectivization. The exploratory analysis of these factors in the corpus data reveals diverse results. The syntactic factor, the adverb of measure and degree *očen′* 'very', which is normally used with adjectives, also combines with participles, and is strongly associated with semantic classes of their base verbs. Nonetheless, the use of *očen′* with a participle only indicates ambiguity when other syntactic factors of adjectivization are in place. The lexical frequency (including the ranks of base verbs and the ratios of participles to other verbal forms) and several morphological types of participles strongly predict ambiguity. Furthermore, past passive and transitive perfective participles not only have the highest mean ratios among the other morphological types of participles, but are also strong predictors of ambiguity.

The linguistic model using weighted syntactic rules shows the highest accuracy in disambiguation compared to the models with weighted morphological rules or the rule based on weights only. All of the syntactic, morphological, and weighted rules combined show the best performance results. Weights are the most effective for removing residual ambiguity (similar to the statistical baseline model), but are outperformed by the models that use factors of adjectivization as constraints.

Acknowledgments

I would describe my PhD journey as a sequence of events in different places that have given me a variety of experiences, knowledge, and a clearer understanding of the academic environment. I am completing this work in a setting that is still unusual to me: on the Norwegian island of Tromsøya, to the north of the Arctic Circle, in the approaching darkness of the polar night, during the COVID-19 pandemic – in short, something I will remember decades later.

I wish to express my gratitude to my supervisors, Svetlana Sokolova, Trond Trosterud, and Eystein Dahl, for their helpful guidance, advice and lively discussions. I appreciate the freedom and space you have given me in my work, and I have certainly gained a lot from this independence in the research process. I am also grateful to Eystein for extensive support in the final stages of my work on the dissertation.

I would like to thank Francis Tyers (Indiana University, USA) for sharing the Apertium annotation tools and giving extremely comprehensive feedback as a final reader. I owe my gratitude to Robert Reynolds (Brigham Young University, USA) for his great assistance at the beginning of my doctoral studies, including informative tutorials on the Russian finite-state tools and help with scripting. Eckhard Bick (University of Southern Denmark) deserves special thanks for discussing the possibilities of Constraint Grammar and for suggesting useful solutions to related questions. I am also grateful to Sjur Mochagen for helping to resolve possible and impossible issues related to the finite-state analyzer and data processing. Special thanks are due to Linda Wiechetek for her advice on multiple occasions, as well as to Lene Antonsen and Ciprian Gerstenberger for offering insightful solutions to my technical questions.

During my doctoral studies, I was pleased to know people from the UiT administration who were always ready to ease my life as a PhD student with relevant advice and support: Beathe Paulsen, Linda Nesby, Kristian Osnes Aambø, and Mayvi B. Johansen. Thank you for consulting with me patiently in any situation and with any question that I had!

This dissertation would not have been possible without the funding provided by the Faculty of Humanities, Social Science and Education at the University of Tromsø, and I am indebted to it for giving me this opportunity. It was a pleasure to be employed at the University of Tromsø and to attend the courses and seminars for PhD students.

In the course of my doctoral studies, I had the wonderful opportunity to have three research stays abroad, during which I came to know many interesting people and projects, and attended enlightening talks, lectures, and seminars. At the Institute of Linguistic Studies (Saint Petersburg, 2017), I was fortunate to work with Sergey Say, who shared invaluable perspectives that are developed further in the dissertation. I am also grateful to Maria Ovsjannikova for her comments on and discussion of the early draft of my dissertation. At the University of Helsinki (2018), I would like to extend my thanks to Jack Rueter, Anssi Yli-Jyrä, and Jörg Tiedemann, who introduced me to the activities of the Language Technology group. I would like to thank Anssi for the outline of weighted disambiguation and crystal clear lectures on neural networks and automata. I am grateful to Yves Scherrer for discussing the design for a statistical analysis. From my research stay at the Saint Petersburg State University in 2018, I am grateful to Elena Yagunova for her interest in and discussion of research topics in my dissertation. I wish to thank Olga Mitrofanova for her insightful feedback on my research, which could have easily taken an alternative path. Finally, I had the opportunity to complete an internship at the Higher School of Economics (Moscow, 2018), and I wish to thank Olga Eremina and Olga Kultepina, among others, for helping me to get started. Many thanks to Galina Kustova (Vinogradov Institute of Russian Language RAS, Moscow) for sharing the database of semantic classes and relevant references. I also thank Vsevolod Kapatsinski (University of Oregon, USA) for his detailed comments on the exploratory analysis of this dissertation.

There are many other people whom I do not mention and to whom I owe acknowledgments. My PhD fellow students from the Breviklia building will be a reminder of fun moments in the doctoral period, thank you for being around!

Most of all, I would like to thank my family and friends who supported me during the entire PhD period, cheered me up, and kept me in touch with the world outside of academia.

Uliana Petrunina
Tromsø, December 2020

Abbreviations

1 first person

3 third person

ACC accusative

ACT active voice

ADJ adjective

ADV adverb(ial)

ANIM animate

DAT dative case

F feminine

GEN genitive

GER gerund

IMP imperative

IND indicative

INF infinitive

INS instrumental

INTERJ interjection

INTR intransitive

LOC locative

M masculine

N noun

NEUT neuter

NOM nominative

NUM numeral

PASS passive voice

PFV perfective

PL plural

POSS possessive

PP past participle

PREP preposition

PRESP present participle

PRO pronoun

PRP prepositional

PRS present

PST past

PTCP participle

REFL reflexive

REL relative

SG singular

SUP superlative

V verb(al)

Note on Transliteration and Glosses

In this dissertation, I transliterate Cyrillic examples using the *International Scholarly System*. For example, a phrase *четвёртый этаж* is transliterated as *četvërtyj ètaž* ‘fourth floor’. The examples that are part of the code in the figures and tables in Chapters 2 and 5 retain their original spelling with optional transliteration; for example, *четвёртый этаж* /četvërtyj ètaž/ ‘fourth floor’.

The letters in the Russian alphabet and their transliterated version using the Scholarly System are presented in Table 1.

Cyrillic	Scholarly
а	*a*
б	*b*
в	*v*
г	*g*
д	*d*
е	*e*
ё	*ë*
ж	*ž*
з	*z*
и	*i*
й	*j*
к	*k*
л	*l*
м	*m*
н	*n*
о	*o*
п	*p*
р	*r*
с	*s*
т	*t*
у	*u*
ф	*f*
х	*x*
ц	*c*
ч	*č*
ш	*š*
щ	*šč*
ъ	″
ы	*y*
ь	′
э	*è*
ю	*ju*
я	*ja*

Table 1: Correspondence table for Cyrillic and transliterated letters.

The annotation of the interlinear glosses complies with the *Leipzig Glossing Rules*. The colon “:” used with a base form specifies a part of speech, and the period “.” specifies a morphosyntactic feature. An example of the gloss for the singular numeral *četvërtyj* ‘fourth’ is presented below.

četvërtyj	*ètaž*
fourth:NUM.SG.M	floor

‘the fourth floor’

In examples without glosses that require the clarification of grammatical categories, I have used the dash “-” to mark a part of speech, and “.” to specify its morphosyntactic feature; for example, *četvërtyj ètaž* ‘fourth-NUM.SG floor’.

Chapter 1

Introduction

1 Motivation

This dissertation presents a comprehensive study of adjectivization in Russian, a derivational process whereby adjectives are formed from participles without affixation. A lack of distinction between participles and adjectives (that is, identical forms) leads to part-of-speech (POS) ambiguity with regard to related morphological forms and meanings. A typical example of such ambiguity is the word form *blestjaščij* 'shining/brilliant', which can be an adjective, as in *blestjaščij pisatel′* 'brilliant writer', or a participle, as in *murav′i, blestjaščie na solnce* 'the ants shining in the sun'. The adjectival form *blestjaščij* 'brilliant' that is homonymous with the participle *blestjaščij* is the end result of adjectivization. The examples demonstrate two essential aspects of adjectivization from both the theoretical and the natural language processing (NLP) perspectives. First, adjectivization is possible due to the ambivalent nature of participles, as they have a verbal syntactic function while conveying an attributive meaning and share the same morphological expression with adjectives. Second, ambiguous participles constitute a source of POS homonymy, which is a common problem managed by various rule-based and statistical methods in automatic text processing. Therefore, I focus on the properties of adjectivized participles, among other aspects of adjectivization, and validate them by comparing them to the evidence in the corpus data. I also extend the scope of the research to the development of a disambiguation model that resolves the homonymy using rules based on the formal distinctions between participles and adjectives.

Given these considerations, the ultimate objective of the dissertation is to explore the factors that constrain adjectivization in Russian and to investigate their relationship with the ambiguity of participles in the empirical data. To do so, I devise a theoretical foundation for adjectivization in Russian that is differentiated at the levels of syntax, morphology, and semantics. Based on these levels, I discern the properties of adjectivized participles, referred to henceforth as factors of adjectivization, which account for cause or result in adjectivization. The research questions relating to this objective are as follows:

- How can one differentiate between adjectivized and unambiguous participles?
- What are the settings in which adjectivization takes place? What are the mechanisms/factors that underlie adjectivization?

An additional objective is to construct a rule-based language model that uses these factors and corpus frequencies to distinguish homonymous participles from adjectives. Language modeling involves the tasks of morphological annotation and the disambiguation of ambiguous participles. The main focus is on the design of the rules that specify the syntactic context and the morphological properties of participles. Implementing corpus frequencies as weights in the models is another important aim within the task of disambiguation. In the framework of adjectivization, the disambiguation experiment clarifies the following questions: How do the weights implemented in the disambiguation model manage ambiguity? Which factors, syntactic or morphological, are best for removing ambiguity? What would be the optimal setting for resolving ambiguity: using weights, or syntactic or morphological factors of adjectivization (jointly or separately)?

2 Scope

In the dissertation, I restricted my study to three domains, namely the theoretical framework of adjectivization, an exploratory analysis of the factors of adjectivization, and the development of the disambiguation model.

The first domain concerns investigating the mechanisms underlying adjectivization and the properties of adjectivized participles from the synchronic perspective. I first outline the general notions of lexical ambiguity, homonymy, and conversion to provide a broader context of adjectivization. I then focus on two main approaches to adjectivization that complement each other: one covers the syntactic context, and the other the internal morphological and semantic properties of participles.

In the second domain, I explore the factors of adjectivization found in corpus data, and test the significance of their relationship to the ambiguity of participles using statistical methods. The study is complemented by the qualitative analysis of actual cases of participles and their contexts drawn from the texts in the corpora. The claims and assumptions relating to the factors of adjectivization are verified by comparing them to the actual evidence taken from the corpus data.

The third domain of the dissertation is an NLP solution for resolving the ambiguity of participles. More specifically, I developed a language model that distinguishes participles from adjectives using the methodological paradigm of Constraint Grammar (CG). I also compared it to an existing statistical solution in the framework of machine learning. The CG method is rule-based, and involves (a) weighting the morphological analyzer for Russian with corpus frequencies of participles and adjectives, and (b) designing CG rules that describe the syntactic and morphological factors of adjectivization.

Consequently, the dissertation has a pronounced quantitative focus on adjectivization, as it tests the theoretical evidence in comparison to the findings in the corpora, and uses it as the basis for the disambiguation model. Both the exploratory study and the disambiguation stem from the theoretical domain because they use the factors of adjectivization as the object of analysis.

3 Background

Adjectivization is investigated in the larger picture of lexical ambiguity and conversion discussed by Manova (2011); Dressler (2005); Lieber (2005); Valera (2015, 2014), among others. In the studies, adjectivization in Russian is approached by describing how (a) adjectivized participles are identified in syntactic contexts, and (b) how their morphosemantic profile can cause them to lose verbal and acquire adjectival properties. The first approach distinguishes between adjectivized and non-adjectivized participles on the basis of the constituents of the immediate context (Say, 2016; Timberlake, 2004). The immediate context may include verbal complements, adjuncts, adverbial modifiers (such as the temporal adverb *davno* 'long ago'), and word order, such as a postposed position to a head noun (that is, a participle preceding its head noun). The lack of these constituents signals that a participle has been adjectivized, in addition to the presence of adverbs of measure and degree (for example, *očen′* 'very'), and adverbs and adjectives of comparative/superlative degree (such as *bolee* 'more' and *samyj* 'the most'). An adjectivized participle can exhibit more adjectival and fewer verbal properties to a greater or lesser extent across the entire continuum of adjectivization. At one extreme of this continuum is an adjectivized participle with an extended and optionally idiomatized meaning. The second approach (b) focuses on the effect of the grammatical meanings and semantics of participles and their corresponding base verbs (Kustova, 2012; Kalakuckaja, 1971; Černega, 2009). These properties, as well as the morphological types of participles (such as past passive or present active participles), favor or disfavor the development of adjectival meaning and the loss of verbal properties in a participle. The question of whether the syntactic behavior of participles reflects the results of adjectivization or causes it, from a synchronic perspective, is also considered in the dissertation.

A rule-based disambiguation model is developed and optimized using CG, which is a language-independent formalism that applies to POS tagging and the shallow parsing of running text using grammatical rules (Tapanainen, 1996; Voutilainen and Tapanainen, 1993; Lindberg and Eineborg, 1998; Karlsson, 1990; Bick, 2000). In the task of tagging, these rules refer to specific morphological features and POSs, are hand written by experts, and can also express surface-syntactic relations (Voutilainen and Tapanainen, 1993). The disambiguation model that I developed consists of the rules employing grammatical categories and morphological features to describe the syntactic context surrounding an ambiguous word form and its internal properties. One of the components of the model was developed using a method discussed by Lindén et al. (2009b,a) and recently adopted by Keleg et al. (2020). The method uses corpus frequencies as unigram probabilities, and compound penalties to disambiguate compound segmentations in Finnish. By analogy with this method, I developed a CG model with a weighted component, which uses the weights to distinguish participles from adjectives. As will be shown later in the dissertation, the results of the model's performance indicated how well the syntactic or morphological factors, or both, differentiated between adjectives and participles, and how weighted parameters managed the task.

4 Relevance

Despite the fact that adjectivization in Russian has been discussed in a number of studies, to my knowledge, there have not yet been approaches that would both present the phenomenon in detail and investigate the factors of adjectivization in the empirical data. This dissertation presents a framework that links two visibly different syntactic and morphosemantic approaches. It singles out the synchronic micro-aspects of adjectivization as factors, and identifies their relevance for favoring or obstructing adjectivization (cause) or signaling its occurrence (result).

The most obvious contribution is in the empirical assessment of the factors in the corpus data. Apart from observations of the frequency distributions of the factors, I statistically tested the strength of the associations, and the significance of the relationships between the factors and the ambiguity of participles. Furthermore, the corpus frequency of the base verbs and the number of participles they form was an additional factor that proved to be a strong predictor of adjectivization. All the quantitative experiments were conducted in parallel with a qualitative analysis of participles and their contexts, as reflected in the distributions.

A practical contribution of the dissertation is the development of a disambiguation model that differentiates between adjectives and participles with a high degree of accuracy and a low ambiguity rate. The methodology is novel because it combines frequency-based weights with CG rules, and distinguishes the performances based on syntactic and morphosyntactic rules. All the factors of adjectivization are formalized in the CG rules to resolve the ambiguity of participles successfully. The design of the rules relies on coarse- and fine-grained syntactic contexts and morphological properties, optionally combined with weights. A specific contribution is the weighting method that I used to generate weights from the frequency list, and the implementation of these weights in the Russian morphological analyzer.

5 Outline

The remainder of the dissertation consists of five chapters. Chapter 2 provides a methodological overview of the main tools and methods used in the empirical analyses and language modeling. It presents the basic concepts of weighted automata and a morphological transducer for Russian, as well as lexicons. It highlights the features of the CG formalism and describes the structure and components of the Russian CG

Chapter 3 discusses theoretical approaches to adjectivization against the general background of POS ambiguity and conversion. The chapter defines the factors of adjectivization and how they operate in the process of adjectivization with regard to their order and causality. In addition to the theory overview, I present some marginal cases of adjectivization and certain syntactic factors via context analysis.

Chapter 4 reports on quantitative studies of the frequency distributions of the factors defined in Chapter 3. The first experiment tests the strength of association between (a) the construction *očen′* + participles and *očen′* + finite verbs, and (b) the semantic classes of the base verbs. The second experiment assesses the effect of frequency and morphosyntactic factors of adjectivization on the ambiguity of participles. This experiment also touches on the pervasiveness of participles with regard to their ambiguity.

Chapter 5 describes the implementation of weights, the design of the gold standard, and the development of the CG disambiguation model, as well as the evaluation thereof. It highlights the individual components of the model based on weights and morphosyntactic factors. The chapter then presents the results of the evaluation of the CG disambiguation and the results obtained by a machine-learning model. This chapter demonstrates how various combinations of the components in the CG model can improve metrics for disambiguation.

Finally, Chapter 6 summarizes the main discussions and findings in the dissertation. It concludes with the findings that emerged from investigating the factors of adjectivization and the ambiguity of participles discussed in the theoretical approaches and in the empirical data. The chapter reviews the development of the disambiguation models and assesses their performances in terms of resolving the ambiguity of participles. It then discusses the implications that arose from the exploratory analysis of the factors and the evaluation results of the disambiguation models. Lastly, the chapter outlines future directions for the analysis of several aspects of adjectivization, and the extension of CG-based disambiguation to other, neural probabilistic and/or vector-based methods.

Chapter 2

Methodology overview

1 Introduction

In this chapter, I review the main tools and methods applied in the quantitative analysis of adjectivized participles and the disambiguation thereof. The primary objective is to provide a clear methodological background to the linguistic analysis of adjectivized participles by focusing on the essential properties of the methods and the functionality of the tools used in the dissertation. For this reason, I only discuss the methods and tools that were used for morphological annotation and disambiguation during the course of the research. The method of morphological annotation is represented by a morphological analyzer that is used to add the morphological information stored in the affixes and stems of an analyzed word form after a word form has been segmented into a stem and a prefix/suffix. The morphological information includes POSs (for example, verbs, adjectives, nouns, and the like) and their grammatical features (such as person, number, tense, voice, and so forth). The method of disambiguation involves the selection of relevant or the removal of irrelevant morphological readings given by the analyzer in order for a word form to retain the best morphological reading. The disambiguation relies on the design of the constraint-based rules in the framework of CG and on the frequency-based probabilities of the words implemented as weights in the analyzer. The constraint-based rules decrease the number of ambiguous readings by analyzing the morphological information pertaining to an ambiguous word form and the constituents of the context surrounding the word form. When the rules are weighted, they also take the value of the weight assigned to an ambiguous word form into consideration, and select or remove the morphological readings with regard to this value.

The language processing tool for annotating and further lemmatizing the participial, finite infinitival verbal, and adjectival word forms investigated in Chapter 4, Section 3, and Chapter 5 is the morphological analyzer for Russian, which is introduced in Section 2 below. The morphological analyzer is a weighted finite-state transducer that recognizes a set of strings and transduces (or translates) each string into another string. The disambiguation model discussed in Chapter 5 was developed on the basis of the same morphological analyzer, and uses the compiler/parser to implement constraint-based and weighted rules.

Section 2 introduces the basic algebraic concepts of a transducer and a weighted finite-state transducer, as well as weights and their use in semirings. It also provides an overview of the domains in which weighted finite-state traducers are applied. Section 3 describes the application of the morphological transducer and its lexicons in morphological annotations and disambiguation. It highlights the properties of the weights and their implementation in the morphological transducer. Section 4 provides a theoretical background to the CG formalism, as well as an overview of the Russian CG and its major components, including constraint-based rules.

2 Weighted finite-state transducer

2.1 Basic definitions

A *graph* G is a non-empty set of states (also referred to as vertices), together with a (possibly empty) set of unordered pairs of different states of G (e.g., Barnard, 2012). An unordered pair of states is an edge of G. $V(G)$ or V is the set of vertices of G; $E(G)$ or E is the set of edges of G.[1]

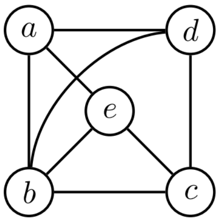


Figure 2.1: An example of an undirected graph G with the vertices a, b, c, d, e.

The graph G illustrated in Figure 2.1 is represented by the set of vertices and edges $G = (V, E)$ where:

- $V = a, b, c, d, e$ is a set of vertices a, b, c, d, e
- $E = \{\{a, b\}; \{b, c\}; \{c, d\}; \{a, d\}; \{a, e\}; \{b, e\}; \{c, d\}\}$ is a set of edges $\{a, b\}$, $\{b, c\}$, $\{c, d\}$.[2]

An *alphabet* is a set A with its elements noted as letters (Pin, 2016; Hanneforth, 2008). A *word* is a finite sequence of the elements in A. The sequence $\{a_0, a_1, ..., a_n\}$ is denoted by the juxtaposition $\{a_0 a_1 ... a_n\}$.

An *automaton*[3] is a directed graph that consists of initial and final states that are distinct from each other, and which are connected consecutively by edges. In a directed graph, the set of edges E is a set of ordered pairs of states of V; that is (u, v), where u and v are labels for given

[1] More precisely, an edge represents a transition from one state to another, drawn as a line in an undirected graph or as an arrow in a directed graph.

[2] These edge definitions are unordered pairs of the vertices a, b, c, d, e.

[3] The description is based on the papers by Yli-Jyrä (2014) and Pin (2016).

states. A set of distinct consecutive transitions with the same states repeated once forms a path. The automaton accepts words via successful computations, and changes its behavior (known as the accepted language) based on the accepted words. An example of a finite automaton over the alphabet $\{a, b\}$ is provided in Figure 2.2. A path in the automaton *A* is a finite sequence of consecutive transitions.

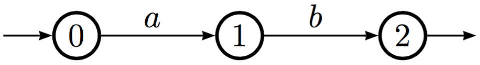


Figure 2.2: An example of a labeled directed graph over the alphabet *A*. The labels are the letters *a* and *b*.

A transducer[4] is an automaton that reads an input word and produces an output. A transducer is deterministic or sequential if and only if each of its states has one transition with any input label, and this label is not an epsilon (empty string). Figure 2.3 illustrates a sequential transducer that has an input string *abaa* transduced into a string of real numbers 01001. The character '|' indicates that a letter *a* or *b* is transduced into a real number 0, 1, or ε from state 1 to state 2, or over state 2 in a loop.

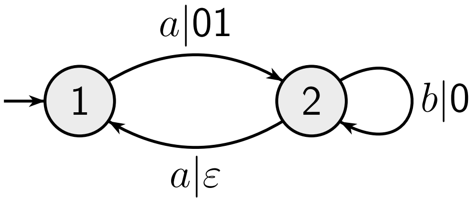


Figure 2.3: A sequential transducer on the input *abaa* and the output 01001 (Pin, 2016). The symbol ε represents an empty string.

A finite-state transducer (FST) is a finite automaton in which state transitions are labeled with both input and output symbols. A path through the transducer encodes a mapping from an input symbol sequence (string) to an output string.

A *weighted finite-state transducer* (WFST) assigns weights on transitions, in addition to the input and output symbols (Mohri et al., 2008). A common set of rational operations such as union, concatenation and Kleene closure is used to combine, optimize, search, and prune weighted transducers (Mohri et al., 2008: 562). The behavior of a weighted transducer can be defined as a function that associates each word with the total weight of its execution. Apart from deciding whether a given word is accepted or not, a weighted transducer also computes the resources, time, cost involved, or the probability of its success when executing the word. Thus, unlike finite transducers, a weighted transducer associates any possible behavior with a weight in addition to the Boolean classifications of "acceptance" and "non-acceptance".

[4] The description is based on the papers by Yli-Jyrä (2014) and Pin (2016).

The *weight* of a transition is a numerical value expressed in real numbers[5] assigned to each of the transitions in a WFST. The weights may encode probabilities, durations, penalties, or any other quantity that accumulates along the paths to compute the overall weight of mapping an input string onto an output string (Mohri et al., 2008). The weight of a sequence of transitions with the initial state and the final state (the global path) is the sum of the weights of each transition.

The WFST in Figure 2.4 is a small-scale finite-state language model (Allauzen et al., 2007).

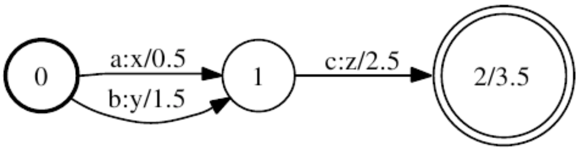


Figure 2.4: An example of a WFST.

In this model, the initial state is labeled 0 and the final state 2. The input labels are *a*, *b*, *c*, and the output labels are *x*, *y*, *z*. Each transition and the final state has a weight associated with it: the transitions $(0, a, x, 1)$[6] and $(0, b,\ y, 1)$ have a weight of 0.5 and 1.5, the transition $(1, c, z, 2)$ has a weight of 2.5., and the final state 2 has a weight of 3.5.[7] The model *transduces* the string *ac* to *xz*, and returns it with the weight of the path as 6.5 $(0.5 + 2.5 + 3.5)$. It also transduces the string *bc* to *yz*, and returns it with the weight of the global path equaling 7.5 $(1.5 + 2.5 + 3.5)$.

The transducer can represent a relationship between two levels of representation; for example, between phones and words by transducing phones into words (Mohri et al., 2008: 562).

A *semiring* is a set with two operations, addition and multiplication, that satisfy axioms such as associativity, commutativity, and distributivity, that are similar to natural numbers with regard to their laws for sums and products (Droste et al., 2009). Weights are part of semiring structures, which are used for computing the weight of the global path using the operations. For example, semirings are used for computing the global path of a sequence of words or letters (based on WFST) with the most probable and highest/lowest weights. Depending on the type of semiring, weights (including weights of the path) may be interpreted as real numbers, probabilities (in the probability semiring), log-probabilities (in the tropical semiring), or costs, Boolean values, strings, distances, feature structure, sets, or matrices. Semirings allow for the definition of a uniform model of weighted transducers for different realizations of weights and their computations.[8]

2.2 Applications of weights and weighted FSTs

Weights are used in a wide range of NLP domains, such as language translation, speech recognition, lexical processing, tagging, summarization, and optical character recognition (OCR; Knight

[5] The weights represent real numbers in this dissertation; however, they may generally refer to any set (Boolean, log, probability, tropical, and so forth).

[6] $(0, a, x, 1)$ represents the transition from state 1 to state 2, wherein 0 and 1 are the labels for the states in the transducer.

[7] *Please Note*: The labels and weights in this automaton are given randomly.

[8] See Appendix A, Table A.1 for the different types of semirings and weights associated with these.

and May 2009). Their use enhances the capabilities of finite-state automata; in other words, by modeling the cost of executing the transition, the amount of resources or time needed for this, or the probability/reliability of its successful execution (Hanneforth, 2008).

In language translation, weights are part of the phrase-for-phrase model devised by Och et al. (1999) and implemented by Kumar and Byrne (2003), as cited in Knight and May (2009). A word-for-word model translating natural language sentences using the WFST in the reverse direction for translating Spanish into English was discussed by Knight and May (2009). Other WFSTs used in machine translation include a hierarchical phrase-based translation system that enables alignment and feature extraction using WFST procedures (de Gispert et al., 2010), and the WFST modeling framework for bitext word alignment and translation (Kumar et al., 2006).

In speech recognition, the chain of transducers and the final language model are weighted by using the method of maximum likelihood and by observing probabilities directly in the available training data and smoothing them. This enables the recovery of the sequence of spoken words that generates a given acoustic speech signal using a standard n-gram model (e.g., Pereira et al., 1994; Knight and May, 2009). Lexical processing tools with weights include the morphological analyses of Turkish and Finnish, in which compounds must often be broken up into separate articles, prepositions, and nouns. Sak et al. (2012) proposed an approach for integrating morphology into an automatic speech recognition (ASR) system for Turkish in the WFST framework as a knowledge source using a morpholexical language model and the lexical transducer of the morphological parser. Smit et al. (2017) implemented subword modeling for a WFST decoder in large-vocabulary continuous speech recognition, including subword segmentation algorithms and ways to mark the word boundaries in subword sequences, and tested it on a variety of Finnish and Estonian datasets.

For tagging, Collins and Singer (1999) constructed an n-gram WFST for modeling POS sequences and a one-state WFST to model substitutions of words by tags, and the classification of named entities in texts using unlabeled examples. Weights are also used for text summarization and headline generation by omitting unnecessary words from an input text and performing a transformation of the remaining words to form an appropriate headline (Zajic et al., 2002). In the domain of OCR, the chain of weighted transducers segments the words into characters, groups the characters into subword sequences, and transforms the sequences into noise-filled sequences (e.g., Knight and May, 2009).

3 Morphological annotation and disambiguation

Morphological annotation can be used separately or prior to disambiguation, and consists of assigning a series of morphological analyses (that is, POSs and morphological properties) to a word form. Disambiguation always follows the annotation, and consists of removing irrelevant readings.

In this dissertation, I used finite-state tools as part of the Helsinki Finite-State Transducer Technology (HFST).[9] As a tool for the annotations in Chapters 4 and 5, I used an HFST-based morphological transducer that allows the implementation of weights.[10] As a formalism for morphological annotation, I used Lexicon Compiler (LexC), a program that reads sets of morphemes and their morphotactic combinations as input, and creates a finite-state transducer of a lexicon as output (Lindén et al., 2009b: 28).

3.1 Morphological transducer

From an end-user perspective, an FST is a data structure that recognizes a set of strings and transduces (or translates) each string into another string. A morphological transducer recognizes words in a given language and produces an analysis of each word. This type of transducer is referred to as a *morphological analyzer*. The analysis usually contains the base form of the word and its POS, followed by morphological information, such as person, gender, number, tense, aspect, mood, voice, degrees of comparison, and so on. A transducer can generate readings (lemmas and morphological information) for a word or words in a sentence. For example, it takes the sentence *на столе стакан* /na stole stakan/ 'there is a glass on the table' as input and outputs the cohorts presented in Figure 2.5. The tags associated with the cohorts are defined in the Russian tag set.[11]

```
"<на>"
"на" Interj
"на" Pr
"<столе>"
"стол" N Msc Inan Sg Loc
"<стакан>"
"стакан" N Msc Inan Sg Nom
"стакан" N Msc Inan Sg Acc
```

Figure 2.5: Morphological analysis of the sentence.

Table 2.1 illustrates a cohort of the morphological analysis of the the word form *на* 'on'. The cohort consists of the word form *на*, base forms or lemmas *на* and their morphological readings as *Interj* (interjection) and *Pr* (preposition), respectively.

The *Interj/Pr* ambiguity is a POS homonymy in which two word forms *на* are graphically identical but share unrelated morphological forms and meanings. For example, with the *Pr* reading, *на* is a preposition denoting a spatial location (for example, *na stole* 'on-PREP the table'),

[9] HFST is a set of software that implements morphological analyzers and tools that are based on weighted and unweighted finite-state transducer technology. HFST is licensed under the GNU Lesser General Public License v3.0. The tools implemented via HFST include morphological analyzers, generators, spell checkers, hyphenators, thesauri, and translation dictionaries, as well as POS taggers.

[10] This is the main distinction between the HFST transducer and the transducer made available as part of Xerox Finite-State Tools (XFST).

[11] Available at: https://giellalt.uit.no/lang/rus/root-morphology.html

Input word form	**Base form**	**Morphological analysis**
<*на*>	*на*	*Interj*
	на	*Pr*

Table 2.1: An example of a cohort that contains the ambiguous base forms *на* 'on/there' used as an interjection and as a preposition.

with the *Interj* reading, *на* is an interjection expressing a strong volition (for example, *na, beri!* 'there-INTERJ, take it!').

A transducer can also be applied in the opposite direction to generate inflected forms from the base form and the morphological information. This type of transducer is referred to as a *generator*.[12] Taking the base form *слово* /slovo/ 'word' and the morphological analysis *слово+N+Neu+Inan+Pl+Acc* (noun, neutral, inanimate, plural, accusative) associated with this word form, the generator produced the following output. As shown in Table 2.2, the generated analysis (2nd column) consists of the base form *слово*, followed by the morphological analysis *+N+Neu+Inan+Pl+Acc*, an inflected accusative plural word form *словá*.

Input analysis	**Output (generated) analysis**
слово+N+Neu+Inan+Pl+Acc	*слово+N+Neu+Inan+Pl+Acc словá*

Table 2.2: Generated morphological analysis for the base form *слово* /slovo/ 'word'.

Morphological transducers often provide multiple analyses per word, and the user must disambiguate the results by choosing the correct analyses. In addition to performing morphological analyses, transducers can function as spell-checkers, translators, and hyphenators.

3.2 Weights in morphological transducers

In morphological transducers based on HFST, weights indicate the probability of a word or its analysis, but may also indicate how well formed the word is. An HFST weight is usually part of the tropical semiring, and is represented as a float; that is, one or more digits that may be preceded by a minus or plus sign, and followed by a comma followed by at least one digit. For example, the regular expression presented in Figure 2.6[13] produces a transducer that maps *abd* onto *acd* with the weight of $0.5 + 0.3 + 0.2 = 1.0$.

```
[ a b:c::0.5 d::0.3 ]::0.2
```

Figure 2.6: An example of a regular expression that produces a transducer that maps the strings *adb* onto *acd* with the weight of 1.0.

[12] See Appendix A, Section 2.

[13] Available at: https://github.com/hfst/hfst/wiki/Weights

This example shows the transition $a : a$ with no weight, followed by the transition $b : c$ with a weight of 0.5, followed by the transition $d : d$ with a weight of 0.3, leading to a final state with a weight of 0.2. If weights are not specified, the HFST tools operate with zero weights.

In the Russian FST, weights are expressed as floats that encode the log-transformed corpus frequencies of adjectival lemmas and verbal lemmas for participles that can accumulate along paths. The FST transforms the lemmas into word forms with the weights associated with the morphological reading of the word form. The weights are then associated with the adjectival reading and participial readings (for example, *PrsAct* as the present active participle, *PstPass* as the past passive participle, and so on) of adjectival and verbal lemmas; the affixes for the respective word forms are unweighted. A weight in the Russian FST is considered to be a penalty; that is, words/analyses with a greater weight are more probable, and greater weights correspond to higher frequencies. When there are several analyses of a word form, they are printed in descending order; in other words, the most probable ones are presented first. However, the default HFST uses an inverted scale whereby greater weights indicate lower frequencies: zero (0) is set as the default weight for known words or words found in the lexicon, and *inf* (infinity) as the weight for unknown words.[14]

Apart from weighting lemmas, as demonstrated in Chapter 5, it is also possible to weight grammatical rules or morphemes, or to use weights to generate word forms. This makes it easier to differentiate among several analyses of a given word in disambiguation.

3.3 Lexicon compiler

The morphological transducer based on HFST writes the morphology in the formalisms LexC (Karttunen, 1994, 1993) and TwolC (Two-Level Compiler). The syntax of LexC and TwolC is written using CG, and files are compiled using *vislcg3* (see Sections 4 and 4.2).

LexC is a high-level programming language used for specifying lexicons. It is based on the two-level morphology enabling the representation of inflectional and derivational morphology in terms of morphophonological phenomena. An *hfst-lexc*[15] (or *lexc*) is a compiler for lexicon definitions written using LexC, which translates the lexicon into the transducer (Beesley and Karttunen, 2003a).

TwolC is a high-level language that is used to describe morphological alternations, such as try:tries, teach:teaching (Beesley and Karttunen, 2003b). Its syntax is based on two-level rules, namely the declarative system of rule constraints proposed by Koskenniemi (1983, 1984). On a practical level, a *twolc* is a compiler for the two-level constraint-based formalism for describing morphophonological and phonological alternations, and other phonological processes (Karttunen et al., 1987; Lindén et al., 2009b).

[14] Sjur Moshagen (personal communication).
[15] Available at: https://github.com/hfst/hfst/wiki/HfstLexc

```
Multichar_Symbols
+Sg +Pl +Prs +Sg3 +N +V +Det

LEXICON Root
verbs ;
nouns ;
determiners ;

LEXICON nouns
flight    Num   "weight: 1.3"      ;
student    Num   "weight: 1.3"      ;
sand Mass   "weight: 1.3" ;
book  Num "weight: 3.3" ;

LEXICON verbs
book    Per "weight: 1.3" ;
read Per "weight: 10.3" ;

LEXICON Num
+N+Sg:    #          ;
+N+Pl:s   #      ;

LEXICON Per
+V+Prs:    #      ;
+V+Prs+Sg3:s   #      ;

LEXICON determiners
the det ;
a det ;

LEXICON det
+Det:    #      ;
```

Figure 2.7: A simple weighted English lexicon.

A lexicon compiled using *hfst-lexc* contains basic information such as stems, inflectional or derivational morphemes, and the appropriate morphological analysis.[16] Figure 2.7 illustrates a basic weighted lexicon compiled by *hfst-lexc*. The example shows that a lexicon defines lexical and grammatical categories as well as weights for each word form that is input for the morphological analysis. *Multichar_Symbols* are tags referring to POSs and morphological properties (*Pl* is plural, *Sg* is singular, *N* is a noun, *A* is an adjective, and *Det* is a determiner). These symbols constitute a morphological reading output by the transducer after the compilation of the lexicon (for example, *student student+N+Sg* 1.299805). *LEXICON Root* refers to the defined POSs: verbs, nouns, and determiners. The other lexicons specify base forms for each POS (for example, *student*, *cat*, *book*, and *the*) and their inflections (such as *+N+Sg* for a singular noun, and *+V+Prs+Sg*3:*s* for adding *s* to a present tense form in the third-person singular). Entries with *"weight*: are allocated in an arbitrary manner, and indicate that a base form can have a corpus probability of 1.3 or 10.3.[17]

The morphological annotation of the sentence *a student reads books* is presented in Table 2.3. The word form *books* is ambiguous, and has two readings: *book+V+Prs+Sg*3 1.299805 for a present tense, third-person singular verbal word form, and *book+N+Pl* 3.299805, for a plural noun.

[16] The documentation is available at: https://github.com/hfst/hfst/wiki/HfstLexc

[17] The values of the weights are defined randomly, and are only for demonstration purposes.

Input word form	Morphological analysis	Weight
a	*a+Det*	0.000000
student	*student+N+Sg*	1.299805
reads	*read+V+Prs+Sg*3	10.299805
books	*book+V+Prs+Sg*3	1.299805
	book+N+Pl	3.299805

Table 2.3: The weighted morphological analysis of the sentence '*a student reads books*'.

3.4 Morphological analyzer and lexicons

The morphological transducer of Russian written by Reynolds (2016; *cf.* Tyers and Reynolds 2015) was used for the morphological annotation described in Chapters 4 and 5. It provides a morphological analysis of word forms in Russian texts by taking such steps as tokenization and the morphological analysis of word forms, and prints the output in a format compatible with CG. The output consists of a word form and a set of base forms with all the possible morphological analyses (POS, tense, aspect, transitivity, gender, case, and so on).

The morphological analysis is made possible by lexicons with base forms and lexicons with affixes classified according to POSs, which are stored at the Giellatekno repository.[18] All the lexicons (including adjectival and verbal lexicons) are based on the resources in Zaliznjak's (2003) dictionary, and can be enhanced manually through the subversion and revision control system.[19] The number of ambiguous readings was sourced from Zaliznjak's (2003) dictionary, and reflects his knowledge and the literature he used to compile the dictionary. For example, *убитый* /ubityj/ 'killed, depressed' receives the reading +*A* as an adjective, and +*V*+*Perf*+*TV*+*PstPss* as a participle based on the entries in Zaliznjak's (2003) dictionary.

A lexicon with base forms consists of base forms (lemmas and roots, or only lemmas), some of which are weighted. Figure 2.8 illustrates a sample of the verbal lexicon.

```
Verb LEXICON
мыть:м нсв_12а_ы́ть "weight: 4.336218" ;
мыться:м нсв_12а_ы́ть_R "weight: 4.122371" ;
```

Figure 2.8: An example of entries from the verbal lexicon for the Russian morphological analyzer.

In the verbal lexicon, *LEXICON Verb*, *мыть* /myt´/ 'wash' is a lemma and *м* is a base (part of a verbal root),[20] while *нсв*_12*а*_*ы́ть* is one of the lexicon types defined in the lexicon with affixes (see the description of the lexicon below). The lemma *мыться* /myt´sja/ 'wash oneself' is a reflexive verb with the same root, as it is related to the affix lexicon *нсв*_12*а*_*ы́ть*_*R* (*R* stands

[18] Available at: https://giellalt.uit.no/

[19] This means that any user authorized to access these lexicons can add new entries of lemmas missing from Zaliznjak's (2003) dictionary but which are found in modern Russian.

[20] The full root is supposed to be *мы*.

for reflexive).

Figure 2.9 shows sample entries from the adjectival lexicon *LEXICON Adjective* that consist of adjectival lemmas (singular, masculine, and nominative) separated from their bases with ":". For example, *государев* /gosudarev/ 'sovereign's' is a lemma, and *госуда́рев* is a base to which an affix is added from a lexicon with affixes; *weight*: 3.790207 is a weight value. The entries presented in Figure 2.9 illustrate five main adjectival types,[21] among others: *государев* /gosudarev/ 'sovereign's' is a possessive adjective, *куриный* /kurinyj/ 'chicken, chicken's' is relative-possessive, *новый* /novyj/ 'new' is qualitative, *следующий* /sledujuščij/ 'following' is pronominal, and *сосновый* /sosnovyj/ 'pine' is relative proper.

```
Adjective LEXICON
государев:госуда́рев п_[мс_1а]ев "weight: 3.790207"
куриный:кури́н п_а "weight: 4.046066" ;
новый:но́в п_а/с' "weight: 5.923538" ;
следующий:сле́дующ п_а "weight: 5.374452" ;
сосновый:сосно́в п_а "weight: 4.226795" ;
```

Figure 2.9: An example of entries from the adjectival lexicon for the Russian morphological analyzer.

A lexicon with affixes (referred to henceforth as an *inflectional lexicon*) consists of a set of lexicons for different types of stems. These lexicons contain morphological readings for these stems and their corresponding affixes, as shown in Figure 2.10.

```
============= Stem type 1 ===============
LEXICON нсв_1а/6а_14_ать_R
:а нсв_1а_R ;
+V+Impf+IV+Prs: PresFut_1_mut_stem_R_Fac ;
+V+Impf+IV:%^M%<у PresAct_R_Fac ;
```

Figure 2.10: An example of entries from the verbal lexicon for the Russian morphological analyzer.

In the verbal inflectional lexicon, the lexicon *LEXICON нсв*_1*а/*6*а*_14_*ать_R* for the type 1 verbal stems (imperfective verbs with the suffix *ать*) contains two types of analyses. The general morphological analysis *+V+Impf+IV* (verb, imperfective, intransitive) points towards the continuation lexicon labeled *PresAc_R_Fac*. The specified analysis *+V+Impf+IV+Prs* (with an additional tag for the present tense *Prs*) points towards a different continuation lexicon labeled *PresFut*_1_*mut_stem_R_Fac*.[22] The transducer identifies lemmas using the lexicon with base forms, and accesses and assigns the morphological properties using the lexicon with affixes. All

[21] For a complete list of adjectival types, see Panova (2010).

[22] The *PresFut*_1_*mut_stem_R_Fac* lexicon points towards the morphological endings in the singular (for example, an inflection *-у-* and a postfix *-сь* as in *+Sg*1*+Fac*:*%ˆM%*<усь *#* ;) and in the plural (for example, a suffix *-ем-* and a postfix *-ся* as in *+Pl*1*+Fac*:*%ˆM%*<емся *#*;).

the stems and affixes, as well as their derivational rules, are based on the information in the digitalized version of Zaliznjak's (2003) dictionary.

In this dissertation, I consider the morphological transducer as a tool for providing a morphological analysis of the adjectivized participles under discussion, and Zaliznjak's (2003) dictionary as a source for this analysis. I understand a morphological analysis to mean a set of morphological properties assigned to a given word form. A morphological analysis provides all the word forms with their morphological readings (one or more). The analysis outputs a cohort that consists of a word form, its base forms, and the readings generated by the analyzer.

The cohort of ambiguous morphological readings for an adjectivized participle output by the morphological analyzer includes two POSs, which are participles and adjectives. Figure 2.11 illustrates the cohort of ambiguous readings for the word form *следующий* /sledujuščij/, which can represent an adjective with the meaning of 'next', or a participle with the meaning of 'following [something]'. The cohort[23] consists of the verbal participial readings *+V*. . . *+PrsAct* with the base form *следовать* /sledovat′/'follow', and adjectival ones *+A* with the base form *следующий* 'next'. The verbal readings has a weight of 4.074219, and the adjectival readings has a weight of 5.592773.

```
следующий следовать+V+Impf+IV+PrsAct+Msc+AnIn+Sg+Nom 4.074219
следующий следовать+V+Impf+IV+PrsAct+Msc+Inan+Sg+Acc 4.074219
следующий следующий+A+Msc+AnIn+Sg+Nom 5.592773
следующий следующий+A+Msc+Inan+Sg+Acc 5.592773
```

Figure 2.11: An example of a cohort for the ambiguous participial word form *следующий* /sledujuščij/ 'next-ADJ, following-PTCP' provided by the Russian morphological analyzer. Information about the POS tags *+V* (verbal forms) and *+A* (adjectival forms) is specified in the verbal and adjectival lexicons of the analyzer. The tag *PrsAct* corresponds to present active participles.

Table 2.4 presents a detailed explanation of one of the lines in the analysis of the word form *следующий*: *следующий следовать+V+Impf+IV+Msc+AnIn+Sg+Nom* 4.074219. The line comprises a word form, a lemma, followed by the POS tag *+V* and tags associated with the morphosyntactic features of the verbal form, such as the imperfective aspect, the present active form, intransitive use, and so forth. A weight of 4.074219 representing the lexical frequency of the verbal lemma is provided at the very end of the line.

[23] The other tags used in the cohort include *Impf* (imperfective aspect), *IV* (intransitive use), *AnIn* (animate, inanimate use), *Inan* (inanimate use), *Sg* (singular), *Nom* (nominative case), *Acc* (accusative case), and *Msc* (masculine gender).

Elements	Definition
следующий	word form
следовать	lemma
V	verb
Impf	imperfective
IV	intransitive
PrsAct	present active
Msc	masculine gender
AnIn	animate or inanimate
Sg	singular
Nom	nominative case
4.074219	weight value

Table 2.4: Elements in the first line of the morphological analysis and their definitions.

The output of double adjectival/participial readings demonstrated in Figure 2.11 is used to refer to the ambiguous participles (as opposed to unambiguous) analyzed in the statistical analysis in Chapter 4 and the disambiguation experiment in Chapter 5. Participles that are assigned verbal readings only (such as *делающие* /delajuščie/ 'doing') by the morphological analyzer are regarded as unambiguous in these studies. Moreover, ambiguous readings as an output of the analyzer represent instances of homonymous participial and adjectival forms in Zaliznjak's (2003) dictionary.

The analysis provided by the transducer makes it possible to explore the morphological properties of ambiguous participles and to estimate their effect on adjectivization. In this way, the transducer (a) solves the task of the morphological analysis of ambiguous word forms, and (b) its output contains lemmas with morphological analyses subject to further exploration and interpretation. For example, I used the transducer to annotate verbal and adjectival lemmas in the exploratory analysis of the morphosyntactic properties of verbal and participial word forms.[24]

4 Constraint Grammar

The disambiguation model (Chapter 5, Section 4) was developed based on annotation using the morphological transducer, and on disambiguation in the framework of the CG formalism (Karlsson, 1990; Voutilainen, 1994). The transducer was run in conjunction with the *vislcg3* parser,[25] which uses CGs for morphological disambiguation, syntactic function assignment, and dependency assignment. A CG for morphological disambiguation, introduced by Karlsson (1990), consists of rules that select or remove a morphological analysis of a word form dependent on its context and/or morphological properties. The weights implemented in the analyzer are used as part of the contexts in the CG rules in combination with the rules that rely on linguistic

[24] See Chapter 4, Section 3.

[25] Available at: https://visl.sdu.dk/cg3.html, https://giellalt.uit.no/tools/docu-vislcg3.html

description.

4.1 Background to CG formalism

CG is a language-independent formalism that is applied to the surface-oriented, morphology-based parsing of unrestricted texts. It is used for disambiguating morphological readings and syntactic labels. All the relevant structures are assigned directly via lexicon, morphology, and simple mappings from morphology onto syntax. The constraints discard as many alternatives as possible to obtain a fully disambiguated sentence with one syntactic label for each word as an optimal result (Karlsson, 1990). The syntax of CG is shallow (*cf.* Listenmaa, 2019), and is intended to represent grammatical and lexical semantic properties based on surface generalizations rather than on deep structures.

CG was initially designed for morphologically rich languages such as Finnish and Turkish, but has been extended to many other European languages such as German, Danish, Portuguese, Spanish, English, French, German, Swedish, Norwegian, and Dutch. It also covers lesser-used languages such as Basque, Catalan, Greenlandic, and North Saami, made available by the Giellatekno Center.[26] CG disambiguation does not depend on the size and availability of the gold standard. However, it does rely on the quality of the gold standard, as CG rules are designed to assess all the cases annotated in the gold standard. CG provides effective and easy-to-use means of morphological, syntactic, and semantic analysis (*cf.* Wiechetek, 2018) and disambiguation, as long as the user has a sufficient linguistic background (for example, in semantics, morphology, or syntax) regarding the linguistic phenomena under discussion.

Another reason that CG is a comprehensive framework for language modeling is its transparent mechanisms of annotation and disambiguation, which can be accessed, tracked, and modified by a linguist without advanced knowledge of formal (that is, non-transformational) grammars, including minimalist grammars. CG grammar statements are closer to real text sentences than they are to theoretical grammar-based approaches such as Lexical Functional Grammar (LFG; Dalrymple 2001), Head-Driven Phrase Structure Grammar (HPSG; Müller 2013; Pollard and Sag 1994; Sag 2003), Extended Standard Theory (EST; *cf.* Marcus 1980), Government and Binding Theory (GB; Berwick and Weinberg 1986), Generalized Phrase Structure Grammar (GPSG; Gazdar et al. 1985; Briscoe et al. 1987), or Tree-Adjoining Grammars (TAG; Joshi 1985). The constraints of CG are based on a lower level of theoretical abstraction than are the rules of current formal syntax. CG constraints include a morphosyntactic structure that is close to "traditional syntax", with core parts such as inflection, concord, and order (Karlsson, 1990). The rules of other constraint-based formalisms and theories of syntax apply to more abstract levels of linguistic representation and require a more advanced knowledge of syntactic theories as well as skills in computing.

[26] Available at the GiellaLT infrastructure; developed and used by Divvun and Giellatekno: https://giellalt.uit.no/lang/index.html

As models of formal linguistic analysis, LFG, HPSG, and GPSG seem to be closer to CG than do EST, GB, and TAG because they are constraint-based grammars. All of them are used to provide a more general framework for theoretical analysis (Malmkjær, 2010). To provide two examples, LFG includes constraints (descriptions involving the objects of the theory; constituent structure trees and functional structures), while HPSG is based on "... (i) an explicit, highly structured representation of grammatical categories, encoded as typed feature structures [...]; (ii) a set of descriptive constraints on the modeled categories expressing linguistic generalizations and declaratively characterizing the expressions admitted as part of the natural language." (Levine and Meurers, 2006: 1–2).

As Karlsson (1990) points out, compared to probabilistic parsers, CG is language-independent, which makes it easy to examine the mechanisms of the formalisms and modify them manually depending on the task, including error diagnosis and the improvement of the probabilistic system. It is also independent of the availability and size of the reference corpus, unlike parsers based on statistical algorithms and neural networks.

CG also has a number of advantages in disambiguation, such as simplicity, well-understood properties, and speed (Koskenniemi, 1990). As its parsing is reductionist, ambiguity is reduced in a gradual and controlled manner, starting from all possible analyses and discarding as many alternative readings as possible. CG can also incorporate probabilistic information for solving smaller numbers of ambiguities at the explicit request of the user after the optimal grammar-based constraints have been applied.

The kernel of CG is the lexicon and the morphological analyzer. CG disambiguation consists of the reduction of existing morphological and syntactic ambiguities using grammar rules, which are referred to as *constraints*. The constraints are interwoven with probabilistic modules/heuristic[27] metrics for solving ambiguities, followed by a fitting procedure for managing parsing failure:

1. optimal linguistic constraints
2. (eventually) more heuristic constraints if the optimal constraints fail
3. reapplication of the optimal constraints in the CG

CG parsing consists of preprocessing, lexicon updating (checking for new entries in the root lexicon), morphological analysis, and syntactic parsing, including morphological disambiguation (*cf.* Karlsson 1995).

4.2 Russian CG

In the framework of disambiguation, I employed the Russian CG (Tyers and Reynolds, 2015; Reynolds, 2016), which is implemented via CG-3 *vislcg3* compiler/parser (*cf.* Bick and Didriksen, 2015; Didriksen and ApS, 2007) as part of the VISL project.[28] The Russian CG and CG-3 *vislcg3*

[27] A heuristic approximates the exact solution in a reasonable time frame.

[28] A research and development project at the Institute of Language and Communication (ISK), University of Southern Denmark, available at: https://visl.sdu.dk/visl2/

compiler/parser are available at the Giellatekno repository.[29]

The file for the Russian CG consists of four sections (according to Bick's (2009) definitions) that are placed in sequential order:

1. The *DELIMITERS* section that defines sentence boundaries, such as *DELIMITERS = "<.>" "<..>" "<...>" "<!>" "<?>" "<¶>"* ;
2. *SETS* with list and set definitions
3. The *CONSTRAINTS* section with rules
4. *END*

4.2.1 Lists and sets

LIST is a set based on a sequence of tags separated by spaces. LIST defines POSs, grammatical properties, morphological properties, and punctuation. Figure 2.12 shows the first group of lists defining nouns (*LIST N*) and pronouns (*LIST Pron*), the second group defining transitive and intransitive uses (*LIST TV* and *LIST IV*), and the third group for the base verbs *идти* /idti/ 'go' and *быть* /byt´/ 'be' (*LIST V/Idjot* and *V/Byt*).

```
LIST N =  N ;
LIST Pron =  Pron ;

LIST TV = TV  ;
LIST IV = IV  ;

LIST V/Idjot = ("идти" vblex pres) ("идти" V Pres) ;
LIST V/Byt = ("быть" vblex) ("быть" V) ;
```

Figure 2.12: An example of LIST entries from the Russian CG.

SET is a set based on existing sets of tags separated by union (*OR* or |), concatenation +, subtraction - and some other operators. SET defines the sets of properties, POSs, punctuation, and sentence boundaries combined. Figure 2.13 presents three examples of *SETS*. The set *NOACC* defines any word *WORD* not in an accusative case *Acc*,[30] *Sem/Distance* defines units of measurement such as kilometer or meter, and *NBR* defines any singular *Sg* or plural *Pl* number.

```
SET NOACC = WORD - Acc ;

SET Sem/Distance = ("километр") | ("метр") ;

SET NBR = Sg | Pl ;
```

Figure 2.13: An example of SET entries from the Russian CG.

[29] Available at: https://github.com/giellalt/lang-rus/blob/develop/src/cg3/disambiguator.cg3

[30] The negation in this set is expressed by "-". We use the set *NOACC* when we want to exclude word forms in the accusative case. For example, we would like to ensure that (a) an ambiguous participle is followed by an accusative noun (a complement) and (b) there are no instances of an adjective (ambiguous with a participle) agreeing in the same case with a noun it modifies. To accomplish this, we assign the constraint with *NOACC* to this participle so that there is no overlap of the accusative case between the participle and its complement.

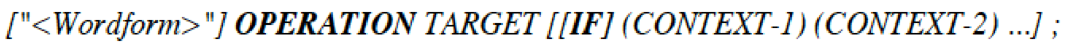
4.2.2 Constraints

The *CONSTRAINT* section only contains rules classified into several types, each of which is delimited by a SECTION statement:[31]

1. Safe (including heuristic) rules
2. Rules for labeling syntax
3. Rules that require correct syntactic function labeling
4. Rules for dealing explicitly with subreadings
5. Customized rules: ‘Test rules for ambiguous participles and adjectives’[32]

The ordering of the rules is relevant because each rule is executed sequentially, and the rules employed in the earlier stages of execution may result in changes in the following stages of disambiguation.

Figure 2.14 illustrates the general form of a CG rule (Bick, 2009: 6).



*["<Wordform>"] **OPERATION** TARGET [[**IF**] (CONTEXT-1) (CONTEXT-2) ...] ;*

Figure 2.14: Elements that constitute a simple CG rule.

Table 2.5 illustrates the elements that specify the CG rule. A contextual test is added each time one wants to specify the context surrounding the cohort, starting from 0 (reference of the cohort), and continuing with 1, 2, 3 (to the right) and/or -1, -2, -3 (to the left).

Element	**Properties**
“<word form>"	a base form
[ ]	an optional element
OPERATION	SELECT (select a reading) or REMOVE (remove a reading) statements
TARGET	a contextual target; that is, a word form which has a tag
IF	an operator of condition (deprecated)
CONTEXT-1, CONTEXT-2	contextual tests that include words surrounding the active cohort

Table 2.5: The elements specifying the CG rule.

Table 2.6 illustrates the elements specifying *CONTEXT*. The contextual test can be extended by adding the globality markers *, ** and 0. One can also specify *CONTEXT* for morphological readings using the operators (*NOT*, *OR*, *LINK*). The statements <*W=MIN*> and <*W=MAX*> in the Russian CG apply to the comparison of weight values.

[31] *Please Note*: The ordering of the rules is a CG3 change to the original CG formalism (Tyers, personal communication).

[32] I wrote this set of rules to disambiguate adjectivized participles.

Element	**Properties**
position markers	self: 0, local right: 1, 2, 3, local left: -1, -2, -3
globality marker	* continue until match, ** continue also across context match, 0* nearest neighbor: search in both directions.
caution marker	C (caution); for example, *1C (only unambiguous readings).
operator	NOT (negation), OR (logical disjunction), LINK (it chains one contextual test to another).
numerical matches	<W=MIN> matching the minimal value, <W=MAX> matching the maximal value. These values are then used for comparison.

Table 2.6: The elements specifying CONTEXT.

4.2.3 Basic examples of rules from the Russian CG

(1) `SELECT Neu IF (0 Dat) (-1 Year) (1 Nazad) ;`

Rule (1) reads: Select the reading neutral gender *Neu* if the word form is in the dative case (0 *Dat*), preceded by the word form of the lexeme *year* (-1 *Year*)[33] and followed by the preposition *nazad* ‘ago’ (1 *Nazad*). For example, for the sentence *Это было год тому назад* ‘It was a year [that-PRO.DAT] ago’, this rule is intended to select the neutral reading for the word form *тому* ‘that’ because this word form is used in the dative case, is preceded by the noun *год* ‘year’, and is followed by the preposition *назад* ‘ago’.

(2) `REMOVE Imper IF (-1C Pr) (0 Imper OR N) (1C Gen);`

Rule (2) reads: Remove the reading of the imperative mood *Imper* if the word form is in the imperative mood or is a noun (0 *Imper OR N*), is preceded by a preposition (-1*C Pr*), and is followed only by a word form in the genitive case (1*C Gen*).

(3) `REMOVE Prp IF (-1C N) (NOT 0 Prop LINK -1 Sem/Person) ;`

Rule (3) reads: Remove the reading with the prepositional case *Prp* if the word form that is not a proper noun *NOT* 0 *Prop* is preceded by a noun denoting a personal title or a profession[34] -1 *Sem/Person*.

Other types of rules used in the section *Customized rules* are weighted rules that include the numerical matches <*W=MAX*> or <*W=MIN*>.

[33] It is part of the temporal entity set `SET Sem/Time = Month | Months | Year | Century | Season | Seasons | TimeOfDay | ("`*период*`") ;`. Year is part of the list `LIST Year = "`*год*`" "`*год*^{1}`" ("`*г.*`" abbr) ;`.

[34] It is part of `SET Sem/Person = PersonTitle | Profession ;`.

(4)
```
SELECT:maxweight (<W=MAX>) IF (0 A)(0 Ptcp);
```

(5)
```
REMOVE:WPTCP-V6.2 V IF (0 A)(NOT 0 Adv)(NOT 0 N)(0 Ptcp + PrsPss + Perf
+ (<W=MIN>)) ;
```

Rule (4) reads: Select the reading with the maximum weight value (<*W=MAX*>) if the word form is an adjective (0 *A*) or a participle (0 *Ptcp*). This rule is placed at the very end of the *Customized rules* section, and targets all the readings that have the largest weight value. Rule (5) reads: Select a verbal reading *V* if a word form is an adjective and not an adverb (0 *A*)(*NOT* 0 *Adv*), and is a past passive perfective participle with the minimum weight value (0 *Ptcp* + *PrsPss* + *Perf* + (<*W=MIN*>)). This rule specifies the target as a word form with a morphological reading *V*, while *SELECT*:*maxweight* specifies the target as a word form with the greatest weight value (<*W=MAX*>). Thus, *SELECT*:*maxweight* may cover many more word forms than *REMOVE*:*WPTCP-V*6.2.

SELECT rules are straightforward: They only retain relevant readings as being disambiguated and discard the rest. *REMOVE* rules are stricter and remove only unwanted readings from the entire cohort, thus leaving some ambiguity, which could be addressed by additional rules in the further stages of disambiguation. For example, after running the rule *SELECT Neu*, only the readings with the neutral gender will be left while, after *REMOVE Imper*, the imperative readings will be removed but all the rest will remain. This is why the safer *REMOVE* rules usually precede the more straightforward *SELECT* ones.

5 Summary

As an abstraction, a WFST computes the weights encoding an estimate (costs, duration, probability, and penalty) for each path that is labeled with a weight. As a tool for linguistic analysis, a morphological transducer recognizes words in a text and provides a morphological analysis of each word. The Russian morphological transducer outputs morphological readings together with the weights for each word. In the Russian transducer, weights are implemented as floats that encode log-transformed corpus frequencies of adjectival and verbal lemmas. Weights and word forms, base forms, lemmas, affixes, and their corresponding morphological readings are presented in the lexicons. The lexicons are compiled using *hfst-lexc* at the moment of the morphological annotation in order for the base and word forms to be output with their respective morphological readings and weights.

The disambiguation is performed by the *viscg3* parser that applies CG for morphological disambiguation and other tasks. The CG formalism is a flexible and multi-purpose formalism that could be applied to various linguistic phenomena in many of the world’s languages. It does not depend on the size of the gold standard, since all the rules are hand written and are tested on smaller samples of corpus examples. In addition, CG rules reduce ambiguity in a controlled manner. It is possible to check and correct rules manually or combine them with a statistical parser.

The file of the Russian CG is divided into delimiters, lists, sets, and constraints (that is, rules). The latter are presented in a sequential order and are classified as safe rules, syntax labeling rules, rules for subreadings, and (also weighted) rules for disambiguating adjectivized participles. Each constraint encodes contexts to describe the context surrounding the words that need to be disambiguated. These contexts can be limited or unlimited, and can be specified with extra conditions. The contexts can also be compared numerically using the weight values.

Chapter 3

Ambiguity and adjectivization

1 Introduction

In this chapter, I develop and discuss the theoretical foundations and key concepts that served as the basis for empirical analysis in my dissertation. I identify specific types of linguistic ambiguity found in Russian, and consider some crucial issues related to the nature of such ambiguity. My investigation centers on the adjectivization of participles (a POS ambiguity) with a specific focus on the syntactic, morphological, and semantic properties of this phenomenon. Adjectivization is a common type of conversion but an atypical way of derivation in Russian because it does not involve any morphological material (suffixes and/or prefixes). It is also a recurrent linguistic phenomenon observed across modern Russian corpus data, and manifests as hybrid syntactic behavior.

(1) a. *neopoznannyj letajuščij ob′ekt*
unidentified flying.NOM.SG.M object
‘unidentified flying ob″ject’

b. *letajuščij za oknami pux*
fly:PRESP outside window fluff
‘[poplar] fluff flying outside the windows’

In (1a), the word form *neopoznannyj* is an adjective modifying the noun *ob″ject* ‘object’, in (1b), it is a participle used with the head-noun *pux* ‘fluff’ and the adjunct *za oknami* ‘outside the windows’. The adjective *letajuščij* is a separate POS that functions as an attribute of nouns, and the corresponding participle is a non-finite verb form (part of the verbal paradigm) that ‘participates’ in the features of both the verb and the adjective (Quirk et al., 1985). The factors differentiating between these ambiguous cases rely on the internal (semantic and morphological) and external (syntactic context) properties of a participial word form. The primary objective of the chapter is to identify which factors pertain to the adjectivization of participles and allow one to differentiate between adjectivized and unambiguous participles. The secondary objective is to examine these factors to determine what causes adjectivization and what results from it.[1]

The chapter is structured as follows. In Section 2, I consider different types of linguistic ambiguity and narrow my focus to POS ambiguity in Russian. In Section 3, I discuss conversion as a universal linguistic process underlying POS ambiguity and focus on adjectivization as one of its types. In Section 4, I first review the major differences between adjectives and participles, and then define the morphosyntactic and semantic properties of adjectivized and unadjectivized participles. More specifically, I outline two major approaches to adjectivization proposed by Say (2016) and Kalakuckaja (1971) and define the morphosyntactic and morphosemantic factors that have been considered to be crucial for adjectivization in these approaches. In Section 4.1, I review the factors of adjectivization based on the syntactic approach that are manifested in the syntactic context based on the syntactic approach. In Section 4.2, I discuss the properties of participles and their corresponding base verbs (such as lexical semantics, grammatical meanings, and so on) identified in the morphosemantic approach by Kalakuckaja (1971); Kolochkova (2011); Kustova (2012), and their role in the restriction or facilitation of adjectivization.[2]

2 Approaches to ambiguity

2.1 Overview of ambiguity

Linguistic ambiguity is the association of a linguistic unit with more than one meaning (Wilson and Keil, 2001). There are two distinct types of ambiguity: (a) lexical ambiguity in which one surface phonetic form has multiple independent lexical representations, and (b) syntactic ambiguity in which one surface string has different underlying syntactic structures (ibid.). Wilson and Keil (ibid.: 14) provide the following examples of these types:

1. Lexical ambiguity:

   (2) *Jeremy went to the **bank**.*

2. Syntactic ambiguity:

   (3) *The company hires **smart women and men**.*

   (4) *The burglar threatened **the student with the knife**.*

In (2), 'bank' can denote both a riverbank and a financial institution. In (3), the sentences can imply that either: (a) the company hires only smart women and men with any level of intelligence,

[1] Assessing the relevance of the factors for resolving ambiguity technically is a more pragmatic objective that will be achieved in Chapter 5.

[2] The importance of some of the factors involved in adjectivization is analyzed empirically in Chapter 4. The factors of adjectivization are also used as the basis for the constraint-based disambiguation models in Chapter 5.

or (b) the company hires smart women and smart men. The example in (4) can imply that (a) the burglar had a knife, or (b) the student had a knife.

More specifically, lexical ambiguity is the association of a word with multiple independent meanings, as in the following examples:

- difference in meaning: *bank* denoting a 'riverbank' and a 'financial institution'
- difference in meaning and syntactic category: *rose* (as a past simple form of a verb and as a noun), *watch* (verb and noun), and *patient* (adjective and noun)

Tanenhaus and Sedivy (2001: 14) compare lexical ambiguity to polysemy ("New York Times" being both a newspaper and the company publishing it) and *vagueness*[3] (the authors provide the example of *cut*, as in "cut the lawn" and "cut the cloth"), but accept that there are fuzzy boundaries between them.

Lexical ambiguity is represented by *polysemy* and *homonymy* (Panman, 1982; Koskela and Murphy, 2006). The studies of homonymy and polysemy by Vinogradov et al. (1960); Matthews (2014); Koskela and Murphy (2006); Ahmanova (1984); Bergenholtz and Agerbo (2014); Derbyshire (1967) allowed me to generalize the authors' claims and assumptions on which I will draw in this chapter.

Polysemy is the association of one and the same lexeme with several etymologically related meanings. It reflects an extension of meanings without changing the grammatical category of a lexeme; an example is the noun *dough*, which denotes 'flour' and 'money'. For this reason, polysemes belong to the same POS and inflectional paradigm (Bergenholtz and Agerbo, 2014). *Homonymy* is a relation between lexemes that are orthographically and/or phonetically identical; for example, the word form *saw* signifies 'a cutting tool' as a noun and is also the past tense form of the verb *see*. While polysemes are grouped under the same entry word, homonyms are viewed as individual dictionary entries (Bergenholtz and Agerbo, 2014). Common homonyms are lexemes that share the same graphical and phonetic forms (full homonymy), although morphemes (affixes) can also be homonymous (partial homonymy). Specific types of homonyms include lexemes with orthographically identical forms (*homographs*) and those with the same pronunciation regardless of the orthography (*homophones*). Homonyms can belong to the same or different word class and morphological paradigm, but differ in semantics; that is, they have etymologically unrelated meanings. In the literature on English, homonyms are distinguished from *homomorphs* by means of morphological and semantic affinity. Homonyms share the same pronunciation and orthography, but have unrelated morphological forms (for example, the noun *saw* 'a tool to split wood' versus the past tense verbal form *saw* '[I] saw'),[4] while homomorphs share the same morphological form but have a different syntactic function (for example, the adjective versus the adverb *fast*). According to Quirk et al.:

[3] *Vagueness* refers to generality and indeterminacy in meaning (Crystal, 2008).

[4] Homonyms can also share a syntactic function; for example, the noun homographs *row* /rau/ 'noisy argument' versus /rəu/ 'line of things, people'.

> There has been considerable disagreement and confusion over the use of the term “homonym”, which has been extended to apply to cases we have referred to as homomorphs. [. . .]We judge red [noun] and red [adjective] to be homomorphs only on the grounds that they share the same stem morpheme, and this in turn rests on the judgment that the two words are related through the process of word formation, in a semantically systematic way (*cf esp* CONVERSION). (Quirk et al., 1985: 70–71)

The instances of the noun *red* and the adjective *red* are characterized by different syntactic functions but share a related meaning, while the typical homonyms *saw* ‘a tool’ and ‘[I] saw’ have unrelated meanings. Studies of homonymy in the Russian structuralist-semantic framework consider homomorphy (*red* used as an adjective and as a noun) to be a particular case of homonymy that is referred to as ‘functional homonymy’ or ‘grammatical homonymy’ (Vinogradov, 1972; Pulkina, 1987) based on the transition/conversion of a lexeme from one word class to another. For example, Babajceva (1967) defines functional homonyms as words that historically were part of the same word class but which came to belong to different classes. Functional homonyms are often associated with the processes of substantivization, adverbalization, and adjectivization.

Quirk et al.’s (1985) approach seems to differentiate homomorphy from polysemy and homonymy based on the etymological relatedness of meanings. I regard such a distinction as unnecessary because I consider homomorphy to be too marginal to be a distinct category of lexical ambiguity. While polysemes are the same lexemes with different meanings, homonyms are distinct lexemes with identical graphical forms and/or pronunciation. This common denominator allows me to regard homomorphs as a type of homonym with a related meaning but a different word class, which also matches the notion of functional homonymy.

2.2 Estimation and classification of ambiguity

In this section, I review the studies that estimated the amount of lexical ambiguity in corpus data, and synthesize different classifications of ambiguity types. First, I investigate how pervasive ambiguity and its specific types are in the corpus data, and the kinds of problems ambiguity creates for tasks related to linguistic analyses. I then specify which linguistic processes, systematic or irregular, underlie specific types of ambiguity. This will provide an overview of lexical ambiguity and underline the relevance of the topic chosen for further investigation in this chapter.

Ambiguity is considered to be one of the central problems in developing language-understanding systems (Allen et al. 1995, as cited in Wilson and Keil 2001). All linguistic input in the processing system contains ambiguous words or sentences, which creates a serious problem for task solving (Afanas′ev and Kobzareva, 2003). In the case of homonymy, a great number of the POS homonyms occur at every level of NLP analysis, and complicate tasks for the intelligent systems of pre-syntactical analysis (ibid.). This is why ambiguity resolution becomes a necessary step in NLP.

Klyšinskij and Rysakov (2015) indicate that homonyms in NLP account for around half of all the words in Russian texts. The researchers identified two main instances of ambiguity in the syntactically tagged corpus of the Russian language (SynTagRus[5]), namely POS ambiguity (24.10%) and POS/morphological ambiguity (51.94%). The former represents ambiguity in the tags standing for POSs only (that is, verbs, nouns, numerals, and so on), whereas the latter comprises both POS tags and the specific morphological readings (such as case, gender, person, and the like) assigned to a word form. Klyšinskij and Rysakov (2015) also observed that most POS and morphological ambiguity resulted from multiple morphosyntactic interpretations.

Tyers and Reynolds (2015) estimated the homography in Russian using text from the Russian version of Wikipedia (see also Reynolds 2016). Their study showed that homographs are a systematic and wide-spread case of morphosyntactic ambiguity leading to the generation of many spurious readings; thus, they regarded it as one of the foundational steps in Russian NLP. Tyers and Reynolds's (2015) defined three types of homographs, namely intra-paradigmatic (90.9% of all the ambiguous tokens), morphosyntactically incongruent (42.7% of all the ambiguous tokens), and morphosyntactically congruent homographs (1.8% of all the ambiguous tokens). Intra-paradigmatic homographs refer to homographic word forms in the same paradigm and the same POS; morphosyntactically incongruent homographs belong to different paradigms and POS, and morphosyntactically congruent homographs belong to different paradigms but the same POS. The proportions of homographs do not add up to 100% because a token may have more than one kind of ambiguity. The cases presented by Tyers and Reynolds belong to the category of lexical ambiguity (including POS ambiguity) and do not involve syntactic structures.

I synthesized Tyers and Reynolds's (2015) types of ambiguity, and the descriptions and examples listed in Ahmanova (1984); Vinogradov et al. (1960); Derbyshire (1967), to form my own classification of ambiguity. The list presented below provides examples of POS ambiguity and two subtypes of inter-paradigmatic ambiguity, namely partial and full ambiguity.[6]

Intra-paradigmatic ambiguity occurs between word forms belonging to the same paradigm and grammatical category (an instance of syncretism); for example, the noun *radosti* 'joy' can be singular genitive, singular dative, singular prepositional, or plural nominative.

POS ambiguity occurs between word forms belonging to different paradigms and POS.

1. Homonyms with unrelated morphological forms and meanings; for example, an imperative verbal form and a possessive pronoun *moj*: 'Wash-V.IMP!' and 'my-PRO.POSS', a verb and a noun *dulo*: 'blow-V' and 'muzzle-N', and a verb and an adverb *počti*: *počti* 'honor-V.IMP' (from *počtit′* 'honor-V') and *počti* 'almost-ADV'.
2. Homonyms with related morphological forms and meanings; for example, homonymous adjective and a noun *moroženoe*: 'ice-ADJ', 'ice-cream-N', a participle and an

[5] Available at: https://github.com/UniversalDependencies/UD_Russian-SynTagRus

[6] For a complete classification of ambiguity, including homonyms, homographs, and polysemes, see Appendix B.1, Table B.1.

adjective *blestjaščij* (from *blestet′* ‘shine-V’) and *blestjaščij* ‘brilliant-ADJ’, and a past tense form and an adjective *smel* (from *smet′* ‘dare-V’) and a short form *smelyj* ‘daring-ADJ’

Inter-paradigmatic ambiguity occurs between word forms belonging to different paradigms but the same POS. Partial ambiguity implies that only certain verbal forms derived from unrelated verbs may be homonymous. In the case of full ambiguity, the entire paradigm is ambiguous.

1. Partial ambiguity between word forms belonging to unrelated verbs; for example, *dobreju*: ‘[I] will finish shaving, am getting kinder’ (from *dobrit′* ‘finish shaving’ and *dobret′* ‘become kinder’), and *leču* ‘I am treating, I am flying’ (from *lečit′* ‘treat, cure’ and *letet′* ‘fly’).
2. Full ambiguity; for example, the imperfective form *zasalivat′* (from *zasalit′* ‘make.greasy-PFV’ and *zasolit′* ‘salt down-PFV’). In this example, the verb *zasalivat′* fully agrees (with the exception of aspect) with the imperfective forms of *zasalit′* and *zasolit′* due to vowel alternation (*o* > *a*) in the verb stem.

Based on the classification, it is clear that POS homonymy differs significantly from intra-paradigmatic and paradigmatic ambiguity. It does not rely on overt morphophonological processes unlike the full paradigmatic ambiguity. It also reflects words with identical graphical forms and pronunciation but with different syntactic functions. The first type (unrelated morphological forms/meanings) appears to be closer to partial paradigmatic homonymy, as both seem to result from the more or less random coincidence of graphical forms with unrelated meanings. The second type[7] (related morphological forms/meanings) does not rely on coincidental matches of graphical forms – it reflects the change in syntactic function without the use of affixes (transposition, conversion) because the morphological paradigm and the semantics of such units maintain affinity. Of all the types presented in the classification, POS homonymy type 2 appears more linguistically engaging – it relies on morphosyntactic and semantic processes that arise from the change in syntactic function. In addition, the notable amount of POS ambiguity found in the corpora (Wikipedia and SynTagRus) may imply that these processes are common in modern Russian. Understanding the phenomena allowed me to develop the theoretical framework for POS ambiguity (homonymy, type 2) in a more systematic way than finding explanations for coincidental intra- and inter-paradigmatic types would have done.

3 Conversion

In this section, I analyze various approaches to conversion, a linguistic process that results in POS homonymy, and define its types and properties. This analysis is important because it will

[7] It is part of the functional homonymy discussed above.

shed further light on the study of adjectivization (as a type of conversion), which is the source of ambiguity for participles. I first provide a detailed account of conversion and clarify the term ‘zero-derivation’, which is often interchangeable with conversion. The properties of both conversion and zero-derivation may overlap fully or partially, and can also represent different aspects of derivation.

3.1 Conversion against the background of word formation in Russian

Petruhina (2006) and Švedova (1980) categorize Russian word formation as being (a) robust due to the limited influence of other languages, and (b) dynamic[8] due to the diversity of its aspects. Diversity is reflected in gaps in word formation chains and in relations of (a)symmetry between formal and semantic structures, among others, as well as the productive[9] models used in speech and more than one possible direction in derivational relations. Finally, word formation is dynamic due to the shift in lexical meaning that derived words undergo.

Words can be formed using one stem (suffixation, prefixation, postfixation, substantivization, and mixed cases) or more than one stem (addition, fusion, abbreviation, and so on; Švedova 1980). Švedova’s (1980) classification also partly overlaps with Arakin’s (2005) types of word formation: *affixation* (Švedova’s (1980) subtypes using affixation from ‘one motivating stem’ and ‘more than one motivating stem’ groups), *compounding* (Švedova’s (1980) fusion and addition), and *affixless means* (Švedova’s (1980) substantivization). From the typological point of view, Arakin (2005) argues that the most productive type of word formation in Russian is affixation due to the “two-morpheme initial structure”,[10] which is also typical of other highly inflected languages such as German and French. By contrast, in English, Chinese, and other languages of the isolating type, the one-morpheme structure is the most popular, and allows words to be converted easily from one POS to another.

The studies that I review in this section indicate that Russian word formation is represented by different types that involve affixation (the most productive type) or exclude it (compounding of affixless means). Although both Arakin Arakin (2005) and Švedova (1980) allude to substantivization as an affixless means of word formation, they do not mention conversion or adjectivization as alternative means.[11] Other studies (e.g., Smirnickij, 1954; Vinogradov, 1954; Panova, 2010) have referred to conversion as a type of derivation and as a cause of homonymy; however, the relationship between conversion and word formation in Russian does not seem to be discussed as frequently as is conversion in English.

[8] The word *dynamic* in this context conveys the meaning of instable, movable because of change, variation, and systemic interaction within a language (*cf.* Wmffre, 2013).

[9] *Productive* and *productivity* in this dissertation refer to a pattern that is used repeatedly in language to produce further instances of the same type (Crystal, 2008). *Productivity* also refers to the notion of generality; namely, the more general a word formation process is, the more productive it will be assumed to be (Katamba, 1993).

[10] This structure consists of a root morpheme and an inflectional morpheme (for simple stems), or root, inflectional and derivational morphemes (for derived stems; Arakin 2005).

[11] Arakin (2005) only refers to conversion as a phenomenon that can be observed in modern English; for example, the lexeme *cold* being used as a noun ‘freezing weather, illness’ or as an adjective ‘chill, wintry’.

3.2 Overview of conversion

3.2.1 Properties and types of conversion

In this section, I review approaches to conversion by considering (a) how it has been defined by different scholars, (b) the properties that characterize it, and (c) the types of which it consists. Conversion is the process of converting a lexeme from one POS to another without affixation. This is a general definition that requires a deeper analysis. Table 3.1 provides a brief summary of the properties inherent in conversion based on my own analysis of each approach proposed by Bauer and Valera (2005a); Valera (2014); Lieber (2005); Greenbaum (1996); Beard (1998); Manova (2011); Dressler and Manova (2002); Dressler (2005); Pšeničnaja (2012).

#	Properties	Examples
P1	change in syntactic function ⇒ change in word-class	Russian *bol′noj*: adjectival function (modifier) *bol′noj škol′nik* ‘an ill-ADJ.M schoolboy’ ⇒ nominal function (subject) *bol′noj zasnul*‘an ill-N.M [person] fell asleep’
P2	affixless means of derivation ⇒ identical graphical and (optional) phonological form of the base and the derived lexemes	*ustavšij* ‘tired-PP [of]’ ⇒ ‘tired-ADJ’
P3	(optional) change in inflectional paradigm	*butcher*-N.SG, *butchers*-N.PL ⇒ *butcher/butcher**s**/butcher**ed**/butcher**ing***-V, no change: *best*-ADJ ⇒ *best*-N ⇒ **best**s***-N.PL
P4	change in semantic properties	English *dressing*-PRESP ‘putting clothes on’ ⇒ *dressing*-N ‘act of dressing, cold sauce for salads, wound solution’
P5	universal process	Russian *blestjaščij*: ‘shining-PRESP’, ‘brilliant-ADJ’, French *étonnant*: ‘surprising-PRESP’, ‘incredible-ADJ’

Table 3.1: Summary of the properties of conversion.

The primary property in the definition of conversion is the *change of word class* of a lexeme (P1), resulting in the *creation of new lexemes* (Bauer and Valera, 2005a; Lieber, 2005; Greenbaum, 1996; Beard, 1998; Manova, 2011). The example *bol′noj* ‘ill’ illustrates the change from the adjectival function of the modifier in *bol′noj škol′nik* ‘an ill schoolboy’ to the nominal function of a subject in *bol′noj zasnul* ‘an ill [patient] fell asleep’.

The change in syntactic function does not affect the graphical form (P2) of the derived lexeme, which is identical to the base one. Phonological changes may parallel the change in the syntactic function without necessarily being motivated by the change in the syntactic function. Identical graphical forms largely explain why conversion results in POS homonymy. The ambiguous word form *ustavšij* ‘tired-PP/ADJ’ is a typical case of conversion that is difficult to disambiguate without considering the syntactic context in which it is used. When looking at the word form *ustavšij* out of context, a native speaker cannot tell if it is an adjective or a participle. The clear-cut distinction can only be ascertained when we consider the context surrounding the ambiguous word form. For example, *ustavšij* is an adjective in the phrase *u nego ustavšij vid* 'he has a weary look/he looks tired' and a participle in *čelovek, ustavšij ot žizni* 'man, tired of life'. The prepositional phrase used as a complement *ot žizni* 'of life' indicates that *ustavšij* is a verbal form. The word form *ustavšij* preposed to the head noun *vid* and used without a complement or adjunct is an adjective qualifying the state of the man.[12]

Another, optional, property of conversion is the *substitution of the inflectional paradigm* (P3; apart from the substitution of syntactic properties and word class values) of the base lexeme with new ones (Dokulil 1968: 225, as cited in Valera 2014). Such substitution results in the difference between the base and the derived lexemes in terms of their inflectional paradigms (morphology),[13] as they belong to different word classes (Valera, 2014: 161). This means that conversion also allows for the addition or omission of inflectional affixes (Bauer and Valera, 2005a). The example of *butcher* shows that, after being converted into a verb, it adopts an inflectional paradigm, as in *butcher**ed**/butcher**ing***. This property is optional because not all converted lexemes show changes in inflection or any inflection at all. To illustrate this point, after conversion, the English lexeme *best*-N does not have any inflection in plural forms (see Table 3.1). Quirk et al. (1985) use this property to distinguish between full (for *butcher*) and partial (for *best*) conversion, while Štekauer et al. (2012) refer to the phenomenon as homonymous and non-homonymous types of conversion. Smirnickij (1954: 12) interprets conversion in Russian as a type of word formation based on the change in the inflectional paradigm, whereby all lexemes share the same paradigm inflections.

Conversion is also related to lexical derivation because lexemes that are converted to different word classes undergo some changes in their meanings (P4; Valera 2014). Dressler and Manova (2002) and Dressler (2005) also argue that conversion is typically characterized by a considerable and regular change in meaning. The example of the lexeme *dressing* illustrates a clear-cut distinction in meaning depending on the grammatical category of the base and the derived lexemes. While the participial meaning corresponds to the process of dressing something/someone, the nominal meaning is extended and signifies the state of dressing, sauce for food, medical treatment for wounds, and so on. One should not confuse semantic change as a property of conversion with what Stein (1977: 229–235, as cited in Štekauer et al. 2012: 214) refers to as *semantic conversion*, which implies semantic change within one lexeme,[14] as in *container* 'magazine, bin' ⇒ *container* 'the contents of the magazine, bin'. The case of container is usually treated as metonymy and as representing the metonymical pattern of the "spatial part and whole" (Peirsman and Geeraerts, 2006: 275–276).

Finally, conversion is found in a number of Indo-European, Afro-Asiatic, Uralic, Niger-Congo,

[12] These, alongside with many other, factors are used for writing CG rules for disambiguation in Chapter 5, Section 4.2.5.1.

[13] Together with their new, functional potential (syntax) and semantics.

[14] This is also a case of polysemy.

and other language families (Štekauer et al., 2012; Valera, 2015: 215–126). Štekauer et al. (ibid.) argue that, in the majority of these languages, conversion applies to the categories of adjective, noun, verb, and adverb, and typically results in pairs such as adjective-verb, noun-adjective, verb-noun, and noun-adjective-verb. Using the examples of Hebrew, Romanian, and Udihe,[15] Štekauer et al. (ibid.: 221–222) maintain that (past and present) participles are a relatively frequent source of conversion in a number of language families. The claim concerning the frequency of conversion is supported by Pšeničnaja (2012), who demonstrates that conversion is a common process in both Russian and French.

No explicit order was specified for these properties in the reviewed literature.[16] Some studies suggest that the change in syntactic function is primary and the overlap of graphical forms secondary (Nikitevič 1985, see also Pšeničnaja 2012; Valera 2015). In addition, Schönefeld (2005) argues that a change in a syntactic function within a word involves a different degree of semantic change. Other studies have suggested the reverse, whereby conversion causes changes in morphological and semantic properties that lead to a change in syntactic function (Vinogradov 1975; Gak 2002, as cited in Pšeničnaja 2012).

3.3 More about morphosyntactic properties of conversion

Schönefeld (2005) and Manova (2011) treat conversion in a broader sense, and assign it additional properties that differ from the properties defined in the previous section.

Schönefeld (2005) argues that conversion[17] is not always considered as a model or as a type of word formation. It can also be a morphosyntactic phenomenon whereby a lexeme in the same word class is used as different morphosyntactic word forms; for example, the English word *tea* can be used as a countable or as an uncountable noun.

Manova's (2011) approach to conversion (in the framework of Natural Morphology[18]) appears to deviate most from the common view of this phenomenon. Instead of covering cases that only concern the change in word class, she suggests the derivational-inflectional continuum for three types of conversion. The first type involves affixation (word class change, word class preservation, and formal conversion). The second type has no affixation (syncretism), and the third type has no affixation, is outside of the continuum, and involves a change in the syntactic role (syntactic conversion).[19] In this way, Manova separates syntax (a change in syntactic function) from derivation (conversion by changing word class) by arguing that (a) derivation (with/without use of affixes) triggers a shift from one word class to another, and (b) a change in syntactic function

[15] Hebrew, Romanian and Udihe belong to the Semitic, Indo-European and Tungusik families, respectively.

[16] These studies do not state explicitly whether the change in a syntactic function occurs diachronically or is viewed from the synchronic perspective.

[17] His definition of conversion is in agreement with Valera (2014), Manova (2005), and Dressler and Manova (2002) with regard to the change in paradigmatic word forms and syntactic functions.

[18] This framework focuses on universal preferences in describing the morphology of a particular language and the ways in which they are related to basic cognitive and semiotic principles (for further discussion, see Manova 2011: 36).

[19] For further details, see Appendix C, Table C.1.

is a syntactic process whereby no paradigmatic change occurs. In English, a morphological (derivational) conversion is illustrated by the lexeme *walk*-N being converted from *walk*-V. These lexemes have different paradigmatic word forms (a noun and a verb), and hence do not agree in gender/number/case. By contrast, the Russian lexemes *sladkoe* 'sweet-ADJ.NEUT' and '[something] sweet-N.NEUT (dessert, third course)' have identical graphical forms; thus, as there are no paradigmatic and morphological changes, there is no link to derivation.

I regard Manova's (2011) approach to be incomplete because she associates syntactic conversion with changes in word class (that is, the input and output of syntactic conversion stand for different word classes), but places it outside of derivation. In addition, she uses the term "conversion" to describe other derivational and inflectional processes not related to conversion, as discussed in the studies given above. Otherwise, her interpretation of syntactic conversion[20] seems to be in line with the concept of conversion outlined and discussed in Section 3.2.1: It is a phenomenon whereby the base lexeme and the derived lexeme share identical graphical forms, but have different syntactic functions and word classes (for example, *sladkoe* 'sweet-ADJ.NEUT' ⇒ '[something] sweet-N.NEUT'; Manova 2011: 112).

3.3.1 Conversion versus zero derivation

A specific aspect of conversion is its relation to another widely used and interchangeable term, 'zero-derivation'. One approach considers these terms to be the same linguistic process (overlap), while the other differentiates between conversion and zero-derivation.

Overlap Manova (2011: 55–57) and Schönefeld (2005) argue that the properties of zero-derivation and conversion could be interchangeable. Schönefeld (2005: 131) terms zero-derivation as the use of a word in one word class as a word in another word class without any formal marker indicating this change, as shown in (5).

(5) noun ⇒ verb *coke, witness, hammer, bomb, stone, shop, bottle, lecture, golf, breakfast* (Hansen et al. 1982: 128ff, as cited in Schönefeld 2005).

This definition is similar to the one suggested by Quirk et al. (1985: 1558), who define conversion as a derivational process whereby a lexeme is converted into a new word class without the addition of an affix. They consider conversion to be analogous with suffixation (for example, *acquit*-V ⇒ *acquitt**al***-N), and interchange the term with 'zero-derivation' because the latter reflects the notion of a 'zero' suffix, which is analogous to actual suffixes.

This and Schönefeld's (2005) definitions create a parallelism between conversion and zero-derivation. Schönefeld (2005) also uses the definition *unmarked change of word category* (or word class) suggested by Vogel (1996: 2, as cited in Schönefeld 2005).

[20] Manova also argues that syntactic conversion resembles syncretism, which implies that the forms that are morphologically identical have different syntactic functions.

Differences First, conversion is related to syntactic processes, while zero-derivation is only part of word formation when a word changes its word class (Marchand, 1969: 360). Marchand uses the term *conversion* to cover specific cases of syntactic transposition (such as *government* in *government job*), which are not related to word formation and derivation, although the term 'lexicalized compound' for such cases is more common (Schlücker, 2019).

Second, zero-derivation involves the alternation between a zero-morpheme and a corresponding overt (or formally expressed) morpheme, as in cash-ø∼ *atom-ize*, in which the zero-morpheme ø in *cash* alternates with the suffix *-ize* in *atomize*. Other examples of the alternation in (6) illustrate that the nominal lexemes *beg-ø* and *lov-ø* are derived with zero-suffixes marked as ø (Manova, 2011: 57):

(6) a. *beg-a-t′* 'run-v' ⇒ *beg-ø* 'run-n'
b. *lov-i-t′* 'catch-v' ⇒ *lov-ø* 'catch-n'

In (6a) and (6b), a zero-morpheme ø in the nominal lexemes *beg-ø* and *lov-ø* 'catch' alternates with the affixes *-a-t′* and *-i-t′* of their base verbs. This alternation reflects the diachronic development of zero-derivation arising from the historical division of the morphological processes into *conversion*, *inflection*, and *zero-derivation*, respectively (Bauer and Valera, 2005b). Zero-derivation in Modern Germanic (and other Indo-European languages) is connected to the originally more explicit morphological system of Indo-European and Germanic, in which a zero-morpheme is used when an overt morpheme is absent (ibid.: 46).

Schönefeld (2005) further refines the definition of conversion and zero-derivation by distinguishing between derivational and syntactic processes. If the unmarked change in word class (defined in *Overlap*) is understood as a change in syntactic function (syntax) within a lexeme (including semantic changes to different degrees), the terms "conversion", "partial conversion", and "transposition" come into play. If this change is understood as a product/process of word formation (derivation: implying semantic change and the emergence of a new, derived word), then either a zero-morpheme or a change in the paradigm will be a marker of this type of word formation. Schönefeld adds that zero-derivation and conversion could also combine within word forms that are graphically identical but conceptually different.

3.4 Summary

As a type of word formation, conversion can be viewed from three different perspectives:

- The derivation of lexemes with/without affixation causing changes in (optionally) morphology, vocabulary and semantics, shared by Lieber (2005), Greenbaum (1996), Beard (1998), Valera (2015, 2014).
- A change in POS with affixation/without affixation (changes in syntax, (optionally) morphology), shared by Marchand (1969); Nikitevič (1985); Kubrjakova (1974); Pšeničnaja

(2012).

- A change in syntactic function leading to changes in syntax, POS, and (optionally) morphology, supported by Manova (2011) and Schönefeld (2005).

The reviewed literature does not state that the properties of conversion follow a clearly fixed order. Pšeničnaja (2012); Vinogradov (1975); Zemskaja (2008); Gak (2002), as cited in Pšeničnaja (2012), only maintain that a change in morphological and semantic properties, triggered by conversion, leads to a change in a syntactic function.

Manova (2011) provides a detailed typology of conversion, but her definition and classification of conversion types appear to be too fluid and broad. This results in some inconsistencies in the systematization of types according to their properties (for example, syntactic conversion is outside of derivation). Apart from the syntactic conversion, the other types are simply means of producing new lexemes from different POS (with affixes) or other grammatical categories, with or without affixes. This is why I am only interested in Manova's (2011) approach to *syntactic conversion*, which corresponds to the general discussion of this phenomenon by other scholars.

These perspectives allowed me to form an appropriate interpretation of conversion in the light of POS ambiguity, type 2 (homonyms with related morphological forms and meanings). Conversion is a derivational process whereby the derived lexeme (a) changes its syntactic function and POS without the use of an affix, (b) shares an identical graphical form with the base lexeme, (c) changes its semantic properties, and (d) can optionally change its inflectional paradigm. The most salient property of conversion is the lack of formal change whereby the meaning is not expressed by morphological material (e.g., Dressler 2005: 269). This process is universal, and is common for adjectives, nouns, verbs, and adverbs.

Studies of zero-derivation have shown that the term *zero-derivation* is understood more generally: It overlaps with conversion or subsumes it. The overlap of zero-derivation and conversion consists of unmarked changes in the word class with a possible change in the syntactic function. The difference lies in morphology: Zero-derivation involves the alternation/replacement of overt morphemes with zero-morphemes and semantic change (Marchand, 1969; Bauer and Valera, 2005b; Schönefeld, 2005), while conversion is considered as a primarily syntax-related process, and can also be viewed outside the realm of derivation and morphology (Schönefeld 2005 and Manova 2011).

I would regard both conversion and zero-derivation to be derivational processes that differ in their use of morphological material. While zero-derivation involves alternation between a zero-morpheme and an overt morpheme, and syntactic change, conversion involves a change in syntactic function without using derivational affixes.

As I am interested in POS homonymy, I focus on the conversion that does cause changes in the inflectional paradigm (P3, Table 3.1) and its specific type, which is adjectivization (see the example of *ustavšij* 'tired-PP/ADJ', ibid.). This phenomenon appears to be complex due to the unmarkedness of ambiguous participials and adjectival word forms, which are difficult to differentiate outside of the syntactic context.

4 Approaches to adjectivization

In the following sections, I focus on adjectivization as a type of conversion whereby participles lose their verbal properties and come to function as adjectives. Manova (2011: 119) interprets adjectivization as a type of syntactic conversion (together with substantivization) whereby participles take the adjectival inflection for gender, number, and case, and can be used as adjectives without any paradigmatic change being involved. (7) shows that the past participle postroennyj agrees in person, number, and case with its head noun *dom* 'house'.

(7) *postroennyj* *dom*
build:PP.PASS.NOM.SG.M house.NOM.SG.M
'a built house'

Say (2016) maintains that a fundamental feature of Russian participles is their hybrid nature: (a) There is no clear-cut line between "participles" and "adjectives", and (b) participles are able to become adjectivized. As they are part of the attributive verbal forms, participles share both verbal and specific adjectival properties. They are created by adding a formant that shapes the participial stem, which is then followed by the inflectional endings of adjectives (Timberlake, 2004). In addition, they can have an attributive function, as adjectives do. Unlike other verbal attributes,[21] participles are productive and can be paraphrased as relative in (8b) and finite verbal clauses (as in (8c); Say 2016). An illustration of this claim is the participial clause *propuskajuščie svet* 'letting the light in' in (8a), which is transformed into a relative clause in (8b) and into a finite verbal construction in (8c).

(8) a. *žaljuzi, propuskajuščie svet*
blinds let:PRESP light
'the blinds letting the light in'

b. *žaljuzi, kotorye propuskajut svet*
blinds that.REL let.PRS.3PL light
'the blinds that let the light in'

c. *žaljuzi propuskajut svet*
blinds let.PRS.3PL light
'the blinds let the light in'

The morphosyntactic properties of participles reflect their verbal behavior (tense, aspect, and verbal arguments), their syntactic function (attributive or predicative), and their compatibility with verbs or adjectives.[22] Since both participles and adjectives share the same morphological

[21] Verbal attributes include adverbial participles (also referred to as and tagged as gerunds); for example, the adverbial participle *čitaja* 'reading', as in *on zasnul, čitaja knigu* 'he fell asleep while reading-GER a book'.

[22] These are adjectives that are not homonymous with participles, as I have not yet begun to discuss the properties

expression and attributive function, it could be challenging to differentiate between them outside of the syntactic context and without considering their semantic properties. Semantic properties are also a significant criterion because the semantics of participles are related directly to the meaning of their base verbs, but can only be loosely bound to the meaning of adjectives by means of synonymy (*cf.* Kustova, 2012).

Questions that may arise based on that which has been stated above concern the main distinctions between adjectivized and non-adjectivized participles in relation to their syntactic, morphological, and semantic properties. How do the morphosyntactic properties of adjectivized participles manifest in the context in which they are used? What linguistic phenomena underlie adjectivization? In this section, I explore these questions by focusing on syntactic and morphosemantic approaches to adjectivization. I also discuss the distinctions among different groups of particles, such as unambiguous and adjectivized participles, and adjectivized participles and deverbal adjectives. Finally, I identify the factors involved in adjectivization that are reflected in the syntactic behavior of adjectivized participles and those that cause adjectivization directly.

4.1 Syntactic approach

In this section, I discuss the study centered on adjectivization as a process that results in a partially or fully adjectivized lexeme. I also classify participles based on their syntactic distinctions with regard to adjectivization, and review the attributive/predicative functions that they may have in a sentence. This approach is aimed at highlighting the distinctions between adjectivized and non-adjectivized participles that account for the syntactic behavior of adjectivized participles in context. The investigation of the syntactic properties of adjectivized participles will allow me to answer the main question: How can one differentiate between ambiguous and unambiguous participles when given the syntactic context in which they are used?

Say's (2016) article presents a bottom-up approach that focuses on the syntactic behavior of adjectivized (that is, ambiguous) participles as opposed to unambiguous ones. The syntactic behavior (or criteria for adjectivization) of the adjectivized participles is illustrated by examples from the disambiguated version of the Russian National Corpus[23] (RNC) that attest to different types of syntactic behavior in a variety of syntactic contexts. Say defined adjectivization as a process of syntactic change and gradual semantic development whereby participles lose the semantic properties they share with finite verbal forms. The process weakens the connection between a participial word form and the verbal paradigm, and leads to the transition of the word form to the class of adjectives.

Adjectivization is viewed as a movement away from verbal properties, leading to a situation of asymmetry, or fewer verbal and more adjectival properties.[24] This implies that, at a point in the 'adjectivization' continuum, a participial lexeme exhibits more adjectival than verbal

of adjectivized participles.

[23] Available at: https://ruscorpora.ru/new/en/index.html.

[24] Say (personal communication).

properties due to syntactic change and a gradual change in the participial semantics. A participle undergoing adjectivization appears to lack (a) the ability to select verbal arguments and specific temporal/spatial modifiers[25] (such as adjuncts), (b) semantic relatedness to its corresponding base verbs, and (c) syntactic paraphrase; that is, relative and finite clauses containing finite verbal forms paraphrased from participial clauses. A participle undergoing adjectivization also tends to be used attributively in a preposed position to a head noun and to combine with adverbs of measure/degree, as well as with adverbs of comparative/superlative degree or superlative adjectives. These properties do not follow any overall chronological order; the only order that can be perceived relates to the loss of verbal properties, starting with syntax and concluding with the semantics of a given participial word form. One property or several properties of adjectivized participles combined can emerge in the process of adjectivization. For example, an adjectivized participle may have no complements (one property), or can be preposed to its head noun without any complements (two properties). Although two or more factors combined may reinforce the loss of verbal properties, the combination is not regarded as a separate criterion for adjectivization, but only as evidence that it can be found in corpus data.

4.1.1 Adjectivization based in syntax

Say (2016) does not clarify whether verbal properties (excluding semantics) are lost in a gradual and/or accumulative manner. An adjectivized lexeme may exhibit one or several signs of adjectivization in an unsystematic manner. Independent factors of adjectivization can be represented by the lack of verbal properties, such as a reduced argument structure or the absence of adjuncts. Other, optional, factors signal that a lexeme is adjectivized if there are also other factors involved. For example, both an adjective and a participle can be preposed to a head noun, but the absence of adjuncts and verbal complements or the presence of a sequence of adjectives describing the syntactic context are more suitable for an adjectivized, rather than an unambiguous, participle. The totality of these properties encompasses a certain stage in the process of adjectivization reflected in the syntactic context. This aspect of Say's (2016) approach appears particularly relevant for the formalization of these properties using CG (see Chapter 5), which is context-based and provides the best disambiguation using the rules describing the syntactic context. For this reason, I focus on each of these properties and examples thereof, as assessed by Say (2016), below. I will also refer to them as *factors of adjectivization* because they contribute to the result; that is, the syntactic behavior of an adjectivized participle in a context. The acceptability judgments for the examples in (9a), (10a)–(10c), and some others are based on the following notation of grammaticality judgments:

- * : an ungrammatical clause
- *?*: the grammaticality of this clause is marginal or dubious

[25] Say (2016) refers to it as a localization of the situation in time and space which is usually conveyed by temporal/spatial adjuncts and adverbial modifiers.

- #: the clause is grammatical overall but unacceptable due to semantic/pragmatic reasons

The lack of ability to select verb arguments[26] is illustrated in (9a) and (10b). In (9a), *rasprostranennaja* 'spread' is used with the agentive complement (i.e. an instrumental noun) *evropejcami* 'by Europeans' in the phrase, which is marginally grammatical compared to *rasprostranennaja bolezn′* 'a wide-spread disease'. In (16a), the direct object *zritelej* 'viewers' in the phrase $^{?}$*potrjasajuščij zritelej fil′m* 'the film amazing the viewers' also appears to be odd or ungrammatical compared to *potrjasajuščij fil′m* 'an amazing film'.

(9) a. $^{?}$*rasprostrannenaja evropejcami bolezn′*
spread:PRESP European.INS.PL decease
'the disease spread by Europeans' (Say, 2016)

b. *potrjasajuščij fil′m*
amazing.NOM.SG.M film
'an amazing film' (ibid.)

c. $^{?}$*potrjasajuščijs zritelej fil′m*
amaze:PRESP viewer.ACC.PL film
'the film amazing the viewers' (ibid.)

The lack of temporal/spatial modification is reflected in the absence of adjuncts of time/place and, allegedly, adverbials[27] used to modify verbal forms. In (10a), the word form *mojuščiesja* used with the noun *oboi* 'wallpaper' is an adjective that modifies the wallpaper as a type of wallpaper made of washable material. The example in (10b) is judged as being semantically unsuitable due to the presence of the adjunct of time *každuju nedelju* 'every week'. Due to the use of the adjunct, the word form *mojuščiesja* no longer conveys the extended adjectival meaning of 'washable', but the verbal action of being washed every week. The phrase *mojuščiesja každuju nedelju oboi* 'the wallpapers washed every week' appears semantically implausible because the adjunct now modifies the word form *mojuščiesja* temporally, and indicates that the wallpapers are used in the same sense as materials that are subject to regular washing. For the same reason, one cannot paraphrase this as a finite or relative clause due to such semantics; that is, wallpapers that are washed every week is an extremely atypical situation in reality. In (10d), the phrase *povyšennye v prošlom godu trebovanija* with the adjunct of time *v prošlom godu* 'last year' is also judged as being semantically unacceptable by Say (2016) because the extended adjectival meaning of 'high' obstructs the use of the adjunct *v prošlom godu*. The judgment in (10d) is not entirely justified because the counterexample in (10e) illustrates that the paraphrase test is valid. Specifically, the participial clause *povyšennye trebovanija* is found in the finite verbal clause *povysili trebovanija* 'increased demands' that is used with an adjunct of time *v ètom godu* 'this year' that conveys the

[26] Verbal arguments include direct objects (intransitive verbs without prepositions in the accusative and genitive cases), indirect objects with/without prepositions in all but nominative cases, the prepositon *po* 'upon' followed by an accusative noun, agentive complement in the instrumental case (for passive participles), and reflexive pronouns (for example, *sebja* 'yourself').

[27] This is my proper claim and is not explicitly stated in Say's (2016) discussion.

temporal modification. This counterexample shows that the meaning of *povyšennye* 'increased' in (10d) is not extended as far as the meaning of *mojuščiesja* 'washable' in (10a) because a paraphrase is possible for *povyšennye trebovanija* and impossible for *mojuščiesja oboi*.

(10) a. *mojuščiesja oboi*
washable.NOM.PL wallpaper
'washable wallpaper' (Say, 2016)

b. #*mojuščiesja každuju nedelju oboi*
wash:PRESP every week wallpaper
'wallpapers washed every week' (ibid.)

c. *povyšennye trebovanija*
high.NOM.PL.NEUT demands
'high demands' (ibid.)

d. #*povyšennye v prošlom godu trebovanija*
increase:PP last year demands
'the demands increased last year' (ibid.)

e. *v ètom godu povysili trebovanija k objazatel'nym rezervam*
this year increase.PST.3PL demand.ACC.PL for obligatory funds
'this year they increased the demands for obligatory funds [of local banks]' (RNC)

The lack of temporal/spatial modification does not necessarily indicate adjectivization, as adjectives can also be used with adjuncts and adverbials. However, when a participial lexeme exhibits this, and several other properties of adjectivization, this factor also contributes to adjectivization. This may imply a certain relationship between the properties of adjectivization, which can be primary (for example, complements) or secondary (for example, adverbials). Other types of adjuncts include phrases headed by prepositions; for example, by the prepositions *v* 'in' and *do* '(up) to', as shown in (11a) and (11b). These cases are clearly related to participles (as they retain prepositions), whereas the word form *povyšennaja* in (11c) can be interpreted as 'high' since it modifies a noun and there is no other verbal (and participial) property that could be related to *povyšennaja*.

(11) a. *cenu, . . . povyšennuju v pol'zu prodavcov-pomeščikov*
price . . . raise:PP in favor landlord-venders
'a price, [artificially] raised in favor of landlord-venders' (RNC)

b. *zarobotnaja plata povyšenaja do 70 rublej*
salary raise:PP to seventy rubles
'salary increased up to 70 rubles' (ibid.)

c. *povyšennaja stavka refinansirovanija*
high.NOM.SG.F rate refinancing
'high refinancing rate' (ibid.)

Adjectivized participles also tend to **combine with adverbs of measure and degree or superlative adjectives**, such as *očen'* 'very', *sliškom* 'overly', and *nastol'ko* 'so' (*cf.* Pulkina, 1987). Say (2016) notes that the participles used with these adverbs are adjectivized if their corresponding base verbs do not combine with the adverbs,[28] and illustrates this property using the examples in (12). The present active participial word form *znajuščij* (from the verb *znat′* 'know') is adjectivized in (12a) because it is used with *očen′*, and its corresponding finite verbal form *znaet* 'knows' in (12b) cannot be used with the adverb.

(12) a. *očen′ znajuščij čelovek*
very knowledgeable.NOM.SG.M person
'a very knowledgeable person'

b. **čelovek očen′ znaet*
person very know.PRS.3SG
'the person knows well'

The example in (13) demonstrates the use of the adjectivized form of the past passive form *produman* 'thought out' with *očen′*. In this example, *produman* is a short form of the adjective *produmannym* 'elaborate', and is used in a row with the synonymous adjective *otčetliv* 'clear'.

(13) *vernuvšis′, rasskazal, čto doklad M. byl očen′ produman, otčetliv*
coming.back told that report M. was very elaborate.INS.SG.M clear.INS.SG.M
'having come back, I said that M.'s report was quite elaborate, clear' (RNC)

Adjectivized participles can form **comparative and superlative degrees**,[29] which does not apply to participles. This property can be illustrated in (14), in which *samyj* 'the most' cannot be combined with the participial word form *zaxvatyvajuščij* 'captivating'.[30]

(14) **samyj zaxvatyvajuščij voobraženie rasskaz*
most.SUP captivate:PRESP imagination story
'the most captivating imagination story'

The preposed position of the adjectivized participle in relation to the head noun is likely to result from a lack of verbal arguments as shown in (15).

(15) *kak éto obyčno delajut kurjaščie ljudi v zadumčivosti*
as it usually do smoke:PRESP people in musing
'as it is usually done by people smoking, lost in thought' (RNC)

Focusing on the syntactic properties, Say (2016) considers the semantics of the adjectivized participle against the background of its syntactic behavior triggered by the change in syntactic function. **The lack of semantic relatedness**[31] between the meaning of participles and their

[28] Say (2016) does not expand further this claim, although the role of gradable meaning or the meaning conveying a value of scale determines the use of verbal and adjectival forms with the adverbs of measure and degree (see Lundquist et al. 2013 and Sičinava 2018).

[29] *Please note*: It is possible to use qualitative adjectives with the adverbs of comparative and superlative degrees or superlative adjectives, while this does not apply to relational adjectives (for example, *stekljanyj* 'glass').

[30] The judgment was checked by nine native speakers of Russian without specializations in philology or linguistics.

[31] Semantic relatedness indicates how close the meaning of a verb form remains to the meaning of a participial form. The meanings of the participle and its base verbs are maximally related when they match.

corresponding base verbs is the result of the idiomatic shift in the lexical meaning, indicating that an adjectivized participle no longer belongs to the verbal paradigm. The shift that occurs in a participial lexeme can also extend its meaning to idiomatic uses; thus, a new adjectival lexeme emerges and becomes a homonym with regard to its corresponding participial lexeme.

(16) a. *blestjaščee vystuplenie*
brilliant.NOM.SG.NEUT performance
'brilliant performance' (Say, 2016)

b. *blestjaščie v temnote limuziny*
gleam:PRESP in darkness limousines
'limousines gleaming in the darkness' (RNC)

The meaning of the word form *blestjaščee* 'brilliant' in (16a) is idiomatic, and deviates significantly from the initial meaning of the base verb *blestet′* 'shine, gleam'. The word form *blestjaščie* 'gleaming' in (16b) conveys a meaning that is closely related to the meaning of *blestet′*. The idiomatic word forms[32] necessarily lack syntactic parallelism, and may share the other syntactic properties discussed above (Say, 2016).

Another example of an extreme semantic shift that leads to the extension of the verbal meaning to a metonymical one is provided in (17b) in contrast to a regular use in (17a).

(17) a. *Tom i ego tovarišči, obižennye blizkimi i roditeljami*
Tom and his mates offend:PP relatives and parents
'Tom and his mates, offended by their relatives and parents' (Say, 2016)

b. *est′ očen′ obižennoe ego pis′mo*
there.is very bitter.NOM.SG.NEUT his letter
'there is a bitter letter from him [Shehtel]' (ibid.)

In (17a), the lexeme *obižennye* 'offended' is a passive past participle formed from the verb *obidet′* 'offend', and is used with the agentive complement *blizkimi* 'relatives'. In (17b), the word form *obižennoe* 'bitter' is devoid of verbal properties due to the absence of verbal complements, and its meaning no longer signifies 'offended [by someone/something]' as in (17a), but 'bitter, expressing feeling of bitterness'. Furthermore, this example involves metonymy, whereby the verbal meaning is extended to modifying the inanimate noun *pis′mo* 'letter' as the subject of the bitter sentiment.

Another property, also illustrated in (17b), is **the lack of syntactic parallelism** between a participial and its corresponding finite verbal/relative clause. Syntactic parallelism relates indirectly to adjectivized participles: It reflects whether acceptability holds for a finite or relative clause paraphrased from a participial clause. A lack of syntactic parallelism indicates that clauses with adjectivized participles cannot be paraphrased as finite or relative clauses. The participial phrase *obižennoe pis′mo* 'bitter letter' cannot be paraphrased as an independent relative clause

[32] Other examples of idiomatic adjectivized participles are *obespečennyj* 'wealthy, not in need, not poor' (from *obespečit′* 'provide with'), *sledujuščij* 'following, next' (from *sledovat′* 'follow'), and *rešajuščij* 'main, major' (from *reshat′* 'decide')

**pis'mo, kotoroe obideli* 'the letter that [they] upset' because the experiential predicate *obideli* does not combine with the semantics of the inanimate noun *pis'mo*.[33] A lack of syntactic parallelism may also be determined by the relationship between the lexical semantics of the adjectivized lexeme and the reduction of its arguments. While the meaning of the base verb *obidet'* 'offend' disallows adding the inanimate noun *pis'mo* 'letter' as an argument, the qualitative meaning of *obižennoe* 'bitter' can modify it.

The retention of adjuncts and verb dependents is the main sign of syntactic verbal behavior that indicates that adjectivization did not occur. It is also reflected in whether a participle is preposed or postposed to a head noun. A preposed participle without complements and/or adjuncts demonstrates adjectival behavior, while a postposed participle exhibits verbal behavior. Thus, the position in relation to a head noun or to a pronoun is not a property in itself, but stems from the retention or loss of verbal arguments in the syntactic context. The presence of arguments reinforces the typical use of a participle in a postposed position, while the absence favors the use of a participle in a preposed position (similar to an adjective). This tendency can be illustrated by the examples in (18).

(18) a. *Kak klass pravjaščij i poraboščënnyj.*
as class rule:PRESP and oppress:PP
'As the class ruling and the oppressed.' (Araneum (Russian, 15.02))

b. *režim islamskogo radikalizma, pravjaščij v Tegerane*
regime Islamic radicalism rule:PRESP in Tehran
'the regime of the Islamic radicalism, ruling in Tehran' (ibid.)

c. *ljudi, pravjaščie mirom, . . . dobilis' svoego položenija*
people rule:PRESP world.INS.SG . . . achieved their status
'and the people, ruling the world, achieved their status for a reason.' (ibid.)

In (18a), the participial form *pravjaščij* 'ruling' has no arguments or adjuncts. Despite the fact that *pravjaščij* is postposed, there are no other formal factors pointing to its verbal properties, which is why it appears to be an adjectivized participle characterizing the social class, together with the other adjectivized participle, *poraboščënnyj* 'oppressed'. In (18b), the form *pravjaščij* is followed by the adjunct of location *v Tegerane* 'in Tehran', which indicates spatial modification, and therefore does not match the criterion of adjectivization. In (18c), the form *pravjaščie* joins the verbal instrumental complement *mirom* 'the world', which illustrates the verbal property of *pravjaščie* and makes it an unambiguous participle.

4.1.2 Marginal cases of adjectivization

Factors of adjectivization such as the lack of temporal/spatial modification, the combination with adverbs of measure/degree, with adverbs of comparative/superlative degree and superlative adjectives, or the use in a preposed position are not absolute, and can also apply to participles that

[33] An inanimate noun refers to a concrete object.

are less adjectivized. These marginal cases are found in specific contexts in which the participle is used, and may arise from the semantic relatedness that adjectivized forms maintain with their respective base verbs. As unambiguous participles share common properties with adjectives by definition, it is hardly surprising that they show some degree of compatibility with the adjectival properties manifested by the factors of adjectivization.

Some unambiguous participles combine with adverbs of measure and degree, as shown in (19). The participle *produman* 'elaborated' is used with both the agentive complement *ego razrabotčikami* 'its developers' and the adverb *nastol'ko* 'so'. For this reason, *produman* displays more verbal than adjectival properties, although it is still used with the adverb *nastol′ko*. The example[34] in (19) demonstrates that the use of participles with an adverb of measure and degree is not a strict criterion for adjectivization, particularly when a participle is followed by a complement.

(19)

funkcional	*. . . nastol′ko*	*produman*	*ego*	*razrabotčikami,*	*čto*
functionality	. . . so	elaborate:PP	its	developers	that

'the functionality is so well elaborated by its developers, that' (RNC)

Similar exceptions apply to the superlative adjective[35] *samyx* 'the most' and the adverb of comparative degree *menee* 'less', as shown in (20).

(20) a.

v	*rejtinge*	*samyx*	*čitaemyx*	*rossijanami*	*knig*
in	rating	most.SUP	read:PRESP	Russians	books

'In the rating of the most read books among Russians' (Vsevoložskie vesti[36])

b.

menee	*odarennyj*	*prirodoj*	*čelovek*	*možet*	*stat′*	*bolee*	*razvitym*
less	gift:PP	nature.INS.SG	person	can	become	more	developed

'a less gifted person can become more developed' (studbooks.net[37])

In (20a), *čitaemyx* 'read' is a present passive participle because it has a dependent noun *rossijanami* 'Russians', it can be transformed into a simple clause with a predicate as a finite synthetic form of the verb, which points towards the verbal paradigm. However, it is still combined with the superlative adjective *samyx* '[the] most'. In (20b), the past participle *odarennij* 'gifted' joins the agentive complement *prirodoj* 'nature' while being preceded by the adverb of comparative degree *menee* 'less' which, according to Say (2016), should decrease its verbal properties. However, the adverb of comparative degree *menee* is used with *odarennyj* in *odarennyj prirodoj* 'gifted by nature' because *odarennyj* conveys the qualitative meaning of 'gifted, talented'. In this case, the factor *menee* seems to be more prominent in showing that a participle is adjectivized, even in the presence of the agentive complement *prirodoj* '[by] nature'.

[34] Say (personal communication) argues that some language users may consider the word form *produman* 'thought out' in *nastol′ko produman ego razrabotčikami* from (19) to be an adjective with an agentive complement. Some language users may also consider it to be a participle used with the adverb of measure and degree *nastol′ko* 'so'. Thus, judgments about the compatibility of the adjectivized word form (for example, *produman*) and the adverb of measure and degree or the use of the agentive complement may differ depending on the speaker.

[35] This term is used in the Reference Grammar of Russian by Timberlake (2004).

[36] Available at: http://vsevvesti.ru/?p=3596

[37] Available at: http://studbooks.net/1928655/pedagogika/zakony_razvitiya

The example in (21a) illustrates that the participle *prodavaemye* 'selling' is adjectivized, given that it has no arguments or local/spatial modifiers, and is preposed to the head noun *produkty* 'foods' while still maintaining semantic affinity with the base verbs.

(21) a. *Samye prodavaemye produkty pitanija – èto produkty social'noj značimosti.*
Most.SUP sell:PRESP products food – this products social importance.
'The best-selling foods are products of social importance.' (business-ideal.ru[38])

b. *samye prodavaemye nam produtkty*
most.SUP sell:PRESP us.DAT.PL products
'the best-selling to us products' (Paraphrased (21a))

c. *samye prodavaemye naseleniju produtkty*
most.SUP sell:PRESP population.DAT.SG products
'the best selling to the population products' (ibid.)

Some examples of exceptions to this case were found by means of consultation with four native speakers of Russian. I asked them to evaluate the sentences in (21b) and (21c). The sentence in (21b) was judged as being unnatural and incorrect by all the speakers; (21c) was judged as being natural by three speakers, and as possibly natural (with some hesitation) by one speaker. This might imply that the grammatical category of the argument has an impact on whether the arguments *nam* '[to] us' and *naseleniju* '[to] the population' will be used with adjectivized participles.

At this stage, the examples above seem to demonstrate that factors of adjectivization operate in an unspecified order whereby a participle loses verbal properties and acquires adjectival ones. Even though there is no consistent order, the interaction of several factors that lead to adjectivization can clearly indicate in the syntactic context that a participle is adjectivized. The semantic shift accompanying the loss of syntactic verbal properties does not necessarily lead to a complete extension of the lexical meaning from a verbal to an attributive one.

4.1.3 Predicativity and attributivity

4.1.3.1 Function and use

The distinction between participles and adjectives is also based on the categories of predicativity and attributivity (e.g., Blox, 1983; Vinogradov, 1954; Žerebilo, 2010; Say, 2016; Parmenova, 2002).

Predicativity is an organizing element of a sentence that conveys the grammatical property of a sentence, and relates this property to the subject via a predicate (a finite verb in a sentence of the verbal type). Predicativity concerns the categories of tense, modality, and person conveyed by a verbal form, or the absence thereof (Vinogradov, 1954). Predicativity has a primary processual meaning (processuality). In the phrase *pojut pticy* 'birds are singing', the syntactic present tense indicates that the utterance takes place at the moment of speech, whereas in *peli pticy* 'birds were

[38] Available at: http://business-ideal.ru/samyj-prodavaemyj-tovar-v-rossiii

singing', the past tense indicates that the utterance took place prior to the moment of speech. The predicative function conveys the main property of the process (processuality). It is attributed to all finite word forms, as well as to the short forms of participles (Žerebilo, 2010).

Attributivity is the property of a grammatical object in a sentence. It is often assigned to adjectives and subordinate clauses that function as modifiers and convey qualitative or relational meanings (Parmenova, 2002). The meaning can convey relativity (for example, *včerašnjaja gazeta* 'yesterday's newspaper'), possessiveness (as in *mamina kniga* 'Mum's book' and *eë kniga* 'her book'), and ordinality (for example, *dom nomer dva* 'house number two' and *vtoroj dom* 'second house').

The attributive function in participles results in gender, number, person, case, and animacy agreement,[39] as shown in (22).

(22) *predprinimateli,* *vooduševlennye* *načatoj oxotoj*
entrepreneur.nom.m.pl.anim inspire:pp.nom.m.pl.anim initiate hunt
'the entrepreneurs, inspired by the hunt started' (RNC)

In (22), the word form *voodučevlennye* 'inspired' is used attributively to modify the noun *predprinimateli* 'entrepreneurs' as being inspired. The necessary agreement in grammatical properties is valid.

Attributive and predicative functions allow us to distinguish between participles and adjectives (*cf.* Parmenova, 2002; Bondarko, 1983; Šaxmatov and Istrina, 1963). Full-form participles function predicatively and attributively,[40] whereas short-form participles only have a predicative function (Šaxmatov and Istrina, 1963: 471). Šaxmatov and Istrina (ibid.: 471) add that full-form present passive participles are rarely used attributively. Short forms of participles are used predicatively in the present tense and the passive voice, as in *on ljubim* 'he is appreciated', and *oni vsemi uvažaemy* 'they are respected by everyone'; examples of short-form past passive participles, with or without linking verbs, are *byl* 'was' and *budet* 'will be' (ibid.).

Participles with a predicative function are used with semi-linking verbs and the linking verb *byt′* 'be' (Letučij, 2018). However, present active participles used predicatively with the copular *byt′* 'be' almost always show a certain degree of adjectivization, for example, *muzej byl potrjasajuščij* 'the museum was amazing', and *izvestie bylo ošelomljajuščim* 'the news was overwhelming' (Say, 2016). These examples also support Timberlake's (2004) claim that active participles are rarely used in constructions in the nominative case. (23) supports his claim: Although the adjectival word form *otjagčajuščimi* 'aggravating' is used predicatively with the copular *byli* 'were', it has no complement or adjunct, and is used in the instrumental (and not in the nominative) case.

[39] See Say (2016) for further discussion.

[40] This applies particularly to participles derived from perfective verbs, as past passive participles derived from imperfective verbs do not have the double *-nn-* and are closer to adjectives.

(23) *Obstojatel′stva prestuplenija byli javno otjagčajuščimi.*
circumstances crime be.PST.3PL obviously aggravating.INS.PL.NEUT.
'The circumstances of the crime were obviously aggravating.' (Timberlake, 2004: 283)

Both unambiguous and adjectivized participles can be used in the attributive position. Unambiguous participles used attributively may not join arguments (including, but not limited to, intransitivity), which is clearly shown by the participles *vspotevšij* 'sweating' and *zadyxajuščijsja* 'panting' in (24).

(24) *kto-to, vspotevšij i zadyxajuščijsja, begaet iz magazina*
someone sweat:PRESP.INTR and pant:PRESP.INTR running from shop
'someone, sweating and out of breath, is running from one shop [to another]' (RNC)

Unambiguous participles can also be preposed to their head noun while being used with complements and adjuncts, as shown in (25). In (25a), the participle *potrjasajuščij* 'amazing' joining the noun object *voobraženie* 'imagination' explicitly shows a verbal behavior, even though it is preposed to the head noun *šedevr* 'masterpiece'. In (25b), the participle *živšuju* 'living' used with the adjunct of place *tam* 'there' demonstrates verbal behavior, despite being used attributively and being preposed to its head noun *doč′* 'daughter'.

(25) a. *potrjasajuščij voobraženie šedevr arxitektury*
amaze:PRESP imagination.ACC.SG masterpiece architecture.
'the masterpiece of architecture, amazing imagination' (Reverso Context: Russian to English[41])

b. *živšuju tam doč′*
live:PRESP there daughter
'his daughter living there' (Timberlake, 2004: 212)

4.1.3.2 Contexts

In the following paragraphs, I review some contexts in which unambiguous and adjectivized participles[42] are used attributively and predicatively. The aim was to observe variation in the syntactic behavior, in addition to the attributive/predicative use in sentences. This variation allows one to observe the factors of adjectivization that can appear in a context with attributive/predicative uses of adjectivized/unambiguous word forms.

The examples in (26) illustrate several contexts in which both adjectivized and unambiguous participles are used attributively. The participles are derived from the verb *sledovat′* 'follow'.

(26) a. *Process vosstanovlenija, verojatno, načnetsja v sledujuščem godu.*
process renovation probably begin next.PRP.SG.M year.
'The process of renovation will probably begin next year.'

[41] Available at: https://goo.gl/Yu2nuh

[42] In this section, I may refer to fully adjectivized participles as adjectives for simplicity of use.

b. *v semnadcatom, no daže i v sledujuščem za nim stoletii*
in seventeenth but even and in follow:PRESP for it decade
'[not only] in the seventeenth century but even in the decade following it' (RNC)

In (26a), the word form *sledujuščem* 'next' denotes something that will occur next or which will take place after something else in a sequence. It is also used without markers of adjectivization (such as complements or adjuncts), and collocates with *godu* 'year', so I would consider it to be an adjective. In (26b), the word form *sledujuščem* 'following' denotes the decade following the period under discussion. The word form is a participle because (a) it conveys a verbal *action*, *the process* of coming next, rather than the state, and (b) it is followed by the prepositional argument *za nim* 'after him'.

The examples in (27) show the use of participial forms derived from the verb *vostrebovat′* 'demand' with attributive and predicative functions. This verb is specific, as only its past passive participles are productive compared to other (including adverbial) participial and finite verbal forms.

(27) a. *antivirusy — naibolee vostrebovannyj segment rynka PO*
anti-virus most popular.NOM.SG.M segment market software
'anti-virus is the most popular segment of the software market' (RNC)

b. *Zaigrannyj sjužetnyj xod okazalsja vostrebovannym.*
worn.out plot device proved demand:PP/ADJ
'The worn-out plot device proved to be demanded/in-demand.' (ibid.)

c. *talant Vallissa okazalsja vostrebovan v šifroval′nom dele*
talent Valissa proved demand:PP in encryption field
'Valissa's talent proved to be demanded in the field of encryption' (ibid.)

d. *informacija okazalas′ vostrebovannoj ne tol′ko častnymi strukturami*
information proved demand:PP not only private organizations.INS.PL
'the information proved to be demanded not only by private organizations' (ibid.)

e. *naibolee vostrebovannoj okazalas′ informacionnaja stojka*
most popular.NOM.SG.F turned.out information desk
'the most in-demand information desk turned out [to be]' (Regnum)[43]

In (27a), the full word form *vostrebovannyj* 'popular, in high demand' is used attributively in a preposed position with the adverbial of the comparative degree *naibolee* 'most' without adjuncts or complements; these syntactic properties thus indicate that *vostrebovannyj* is an adjective. In (27b), the full form *vostrebovannym* 'in-demand/demanded' is used predicatively with the semi-linking verb *okazalsja* 'proved to be' in the postposed position; however, there are no other contextual markers of verbal properties. In (27c), *vostrebovan* 'demanded' is a participle because (a) it is a shortened form of *vostrebovannyj*, (b) it conveys temporality marked by the prepositional phrase (used as an adjunct) *v šifroval′nom dele* 'in the field of encryption', and (c) it is used predicatively with the semi-linking verb *okazalsja* 'turned out'. In (27d), *vostrebovan* 'demanded'

[43] Available at: https://regnum.ru/news/society/2548398.html

is a participial form because it is used predicatively with the semi-linking verb *okazalas′* ‘turned out’ and the complement *častnymi strukturami* ‘private organizations’. In (27e), the full word form *vostrebovannoj* ‘popular’ is used predicatively with the semi-linking verb *okazalas′* ‘turned out’, which conveys the verbal property. However, its use with the adverbial of superlative degree *naibolee* ‘the most’, in a preposed position to a head noun, the absence of adjuncts, and the meaning of ‘in-demand, popular’ qualifying the noun *stojka* ‘desk’, reinforces its adjectival properties; thus, I would consider it to be an adjective.

The observations in (26) and (27) show that, apart from attributive and predicative uses, additional adjectivization criteria are necessary to identify whether a word form is adjectivized. This implies that attributive and predicative uses do not appear to be primary indicators of adjectivization. I apply the predicative uses of short-form past passive participles and full-form present active participles with copula and semi-linking verbs as factors in adjectivization in the CG rules.[44]

4.1.4 Classification of participles

Ambiguous participles and adjectives could differ systematically in their syntactic behavior, both among themselves and compared to unambiguous participles. Say does not classify participles as adjectivized or non-adjectivized, as each of the participial lexemes can manifest more adjectival and less verbal behavior depending on the context. The only distinction that Say makes is between participles and deverbal adjectives. For this reason, I considered it necessary to group participles using the criteria of ambiguous/unambiguous (verbal paradigm only or verbal and/or adjectival paradigm), and idiomatized/non-idiomatized meanings. Instead, I defined these groups as unambiguous participles, deverbal adjectives, and adjectivized (ambiguous) participles.

Unambiguous participles[45] have all the properties attributed to verbal behavior (only the verbal paradigm). For example, the present active participial word form *žalejuščij* ‘having pity’ only belongs to the verbal paradigm, and has no properties that would bring it closer to the adjectival paradigm. These participles are not affected by syntactic conversion and do not undergo adjectivization (Manova, 2011). The rules that enable the morphotactic difference between participles and adjectives are not part of syntactic conversion and adjectivization:

- *-nn-* for long forms of participles and *-n-* for long forms of adjectives; for example, *ranennyj* ‘injured-PP’ versus *ranenyj* ‘injured-ADJ’
- *-n-* for short forms of participles and *-nn-* for short forms of adjectives; for example, *obrazovana* ‘formed-PP versus *obrazovanna* ‘educated-ADJ’
- suffix variation between *-šč-* for participles and *-č-* for adjectives; for example, *sidjaščij* ‘sitting-PRESP’ versus *sidjačij* ‘sedentary-ADJ’

[44] See Chapter 5.

[45] I also refer to them as regular participles throughout the text.

Manova (2011) adds that these rules are not absolute and only hold for a restricted number of past passive and present active participles and their corresponding adjectives.

Unambiguous deverbal adjectives or *pseudoparticiples* are adjectives that were formerly adjectivized from participles, and cannot have corresponding forms in the verbal paradigm (Plungjan, 2010). Plungjan argues that these presudoparticiples are the main source of deverbal adjectives, and arose from Old Russian participles with the suffix *-l-*; that is, the adjectives *zagorelyj* 'tanned' (Plungjan, 2010: 3) and *kislyj* 'acid (flavor)', or from active participles with the suffixes *-uč-/-ač-* that were replaced by Old Church Slavonic forms with the suffixes *-ušč-/-ašč-*; for example, the suffix *-ašč-* being used with the participle *ležačij* 'lying, bed-ridden' versus *ležaščij* 'lying, reclined'(Plungjan, 2010: 3).

Adjectivized participles[46] are the intermediate group between unambiguous participles and unambiguous deverbal adjectives. Their main characteristic is the shared ambiguity in participial and adjectival forms. Adjectivized participles undergo full or partial loss of their verbal properties and the acquisition of adjectival ones; for example, the word form of a present active participle *kurjaščij* 'smoking (a cigarette)' can also be used as an adjective signifying 'smoking (compartment)', and a noun denoting a 'smoker'.[47] This group also includes deverbal adjectives that arose from the participles, which exist synchronically but are individually idiomatized (for example, the adjective *rešajuvsčij* 'decisive' in (28a)) and non-productive passive imperfective participles, such as the adjective *terpimyj* 'tolerant' in (28b).

(28) a. *rešajuvsčij* *udar*
decisive.NOM.SG.F strike
'decisive strike' (Plungjan, 2010: 2)

b. *terpimyj*
tolerant.NOM.SG.M
'tolerant' (ibid.: 2)

These deverbal adjectives are still homonymous with the corresponding unambiguous participial forms from which they were adjectivized. Thus, adjectivized participial forms can either be close to the adjectival paradigm or already be part of it while still belonging to the verbal class, directly or by derivation (*cf.* Say, 2016).

Unambiguous participles are part of the verbal paradigm and the farthest from the adjectival one, while unambiguous deverbal adjectives are the most remote forms from the verbal class and are linked to verbs only diachronically. Ambiguous deverbal adjectives are only connected to the verbal class derivationally (see Say 2016); that is, by sharing the same graphical form resulting from the conversion of a participial lexeme into an adjectival one.

[46] The term *adjectivized* can be synonymous with *ambiguous*, although the ambiguity of a participle does not imply that it has all the necessary properties of adjectivization.

[47] The lexeme *kurjaščij* can also be ambiguous, with its corresponding noun meaning 'smoker' as a case of substantivization.

4.1.5 Summary

This section discusses the general properties of participles, the range of features distinguishing participles from adjectives, the internal (or within class) distinctions between regular and adjectivized participles, and adjectivized participles and deverbal adjectives, and the role of the attributive and predicative functions of participial word forms with regard to their syntactic behavior and properties of adjectivization.

Say (2016) interprets adjectivization as a process that ends with the complete deverbalization of a participle, primarily manifested by the loss of its verbal syntactic properties and the extension of the lexical meaning derived from the base verb. In this way, adjectivization becomes the main source of POS homonymy between participial and adjectival forms. The process is based on a number of properties that reflect the syntactic behavior of the adjectivized participle (factors of adjectivization), namely the weakening of verbal properties (loss of adjuncts and arguments, combination with adverbs of measure and degree) and the gradual divergence from a verbal lexical to an adjectival meaning (loss of semantic relatedness). Say (2016) does not define the order (chronological or else) that these factors follow. The process may begin to operate on some atomic properties and then continue with larger sets of the properties. Different properties of adjectivization can also interact with each other, and this interaction can occur at any point in the process. Say (2016) maintains that only idiomatic forms of adjectivized participles become part of the adjectival paradigm; otherwise, he does not specify whether a non-idiomatic participial form showing the syntactic behavior of an adjective is an adjective.

Say's (2016) approach does not establish clear-cut boundaries between the classes of participles and adjectives due to the continuity of adjectivization. This makes it difficult to distinguish between adjectivized and unambiguous participles, as any participle can be fully or partially adjectivized. Moreover, Say (2016) does not specify whether the shift in lexical meaning affects the syntactic behavior of the adjectivized participles or vice versa, or whether these changes occur simultaneously. In addition, he illustrates and discusses the properties of adjectivization using corpus examples; however, his approach is not based on empirical corpus evidence. The absence of boundaries can become problematic with regard to the resolution of ambiguity and its more experimental analysis. Say (2016) does not highlight the semantics of the base verbs[48] and the morphological properties of participles (tense, voice, and aspect), which might also affect the syntactic behavior of a participle. The gaps mentioned above may have arisen due to the lack of generalization of these properties in Say's (2016) study because he limits his analysis to specific cases from the corpora.

Predicativity and attributivity are grammatical (syntactic) properties of a (verbal or adjectival) word form in a context that can characterize it as a participle or an adjective. Apart from the common view that verbal properties are connected to predicativity and adjectival properties to

[48] For example, the verb *podhodit′* is polysemous, it conveys an action in the meaning of 'come up' and a state in the meaning of 'suit, fit'.

attributivity, the attributive and predicative functions of participles are not as prominent as are the syntactic properties of adjectivization. Participles used predicatively with copular and semi-linking verbs tend to be unambiguous, while those used attributively without adjuncts/complements tend to be adjectivized.

In the classification of participles, I divided them into the three categories of unambiguous, adjectivized participles, and unambiguous deverbal adjectives. While unambiguous and adjectivized participles can be homonymous with each other, unambiguous deverbal adjectives are only part of the adjectival paradigm. Adjectivized participles share the syntactic and semantic properties of adjectives acquired through the process of adjectivization.

4.2 Morphosemantic approach

In this section, I discuss the approach that brings into focus the internal (semantic and grammatical) properties of participles that favor or disfavor adjectivization. This approach, or some of its aspects, was discussed by Bardina (2003); Lopatin (1966); Kustova (2012); Kolochkova (2011); Kalakuckaja (1971); Černega (2009). The prerequisites for adjectivization include (among others) the metaphorical meaning of participles, their phraseological use, and the passive voice (almost devoid of the temporal meaning in participles). Within this approach, adjectivization (also referred to as semantic derivation) consists of the gradual (accumulative) extension of the lexical semantics (e.g., Kolochkova, 2011). This extension results in a qualitative grammatical shift, leading to homonymy between a participial and adjectivized lexeme (which becomes part of the adjectival paradigm).

The common point in the syntactic approach discussed previously and in the morphosemantic approach concerns the main function and effect of adjectivization. Specifically, adjectivization is an affixless derivation (*cf.* conversion) that causes a gradual change in the lexical meaning of a participial lexeme and results in POS homonymy between participles and adjectives.

The difference between the syntactic and morphosemantic approaches lies in the object of their focus; the sequence and nature of the process. Adjectivization within the syntactic approach is based on an unordered process: A participle can be adjectivized at any point in the process depending on the properties of the syntactic context in which it is used. Adjectivization in the morphosemantic approach focuses on the internal properties of participles. Moreover, semantic change is primary and syntactic change is secondary in this approach.

First, I provide an overview of adjectivization against the background of semantic change. I then focus on Kustova's (2012) study of adjectivized participles, and identify the properties that can relate to adjectives. This concerns a synonymous relationship between the meanings of adjectivized participles and adjectives. Second, I assess Kalakuckaja's (1971) classification of types of participles that are inclined or disinclined to be adjectivized by considering their properties of tense, voice, transitivity, and aspect. More specifically, I highlight the semantics of the base verbs that form participles (for example, the semantics of transitive and intransitive verbs), including the ability of verbal meaning to extend and become more abstract rather than concrete (Kolochkova 2011 and Kalakuckaja 1971). This will highlight the role of semantics and the interaction of semantics with morphological properties (that is, tense, voice, aspect, and affixes, among others) in relation to adjectivization.

The main research question that I aim to answer is: Under what conditions can a participle detach from the verbal paradigm and become an adjectival lexeme? More specific questions concern the role of semantic change and morphological properties in adjectivization, and whether they cause or prevent this process.

4.2.1 Adjectivization based in semantics

In this approach, adjectivization is interpreted as a process whereby a participle becomes part of the adjectival paradigm by means of (a) the extension or specification of lexical meaning,[49] and (b) a change of meaning in aspect, tense, and voice. Adjectivization is also approached as the result of this process, often represented by homonymy that occurs after the semantic shift and syntactic change (e.g., Bardina, 2003).

In case (a), a participle shifts semantically from the base verbs; that is, it acquires idiomatic meanings specific to qualitative adjectives, as in *blestjaščie sposobnosti* 'brilliant abilities', *podavlennoe nastroenie* 'subdued mood', *oživlennaja beseda* 'lively conversation', and *upavšij golos* 'dismal voice'. The semantic shift leads to the emergence of the grammatical and lexical derivational properties of qualitative adjectives, such as the ability to be used with adverbs of comparative/superlative degree or with superlative adjectives, as in *bolee cvetuščij vid* 'fresher complexion', and *samyj vydajuščijsja učenyj* 'the most outstanding scientist'. The end result of case (a) is the reinforcement of the lexical meaning (qualitative or relational) intrinsic to adjectives. Adjectivization is complete when an adjectivized participle is used in metonymical or metaphorical meaning (Černega, 2009).

In case (b), adjectivization is based on the functional reanalysis of the meanings of aspect, tense, and voice (without losing any connection with them). For example, past tense perfective participles typically denote a state that is the result of the preceding action; for example, *osveščennye okna* 'lit windows', and *promerzšaja zemlja* 'frozen ground', while active and passive present tense participles indicate the ability to perform an action and being subject to any external action, as in *v′juščiesja rastenija* 'vines', *pečatajuščee ustrojstvo* 'printing device', and *neržavejuščaja stal′* 'stainless steel'. Bardina (2003) also views case (b) as the loss of the temporal property of time.

In the light of the semantic shift and change in verbal meaning, adjectivization is viewed as an individual grammatical property of participles that is not, as a rule, influenced by the syntactic context (Kalakuckaja, 1971).

The role of semantic change in adjectivization has mainly been studied by Kolochkova (2011); Kalakuckaja (1971); Bardina (2003); Černega (2009). In their approach, lexical semantic

[49] The extension often includes idiomatization.

change is primary and leads to obligatory changes in grammatical proprieties. The end result of these changes is the homonymy between participles and adjectives arising from the attributive meaning of the adjectivized participles and their lack of verbal properties. The conditions for adjectivization include a preposed position of participles, the loss of verbal government, and the extension of lexical meaning (Černega, 2009).

Apart from adjectivization, semantic change typically leads to POS homonymy between participles/adjectives and nouns (substantivization), nouns and prepositions (prepositionalization), adverbs,[50] and other grammatical categories (Bardina, 2003: 164). Participles are ambivalent; that is, they combine the grammatical properties of a verb (action, time) and an adjective (attribute of a person or an object; Bardina 2003). Consequently, the verbal properties can weaken and the adjectival ones expand, which makes adjectivization more prominent in participles than in numerals or nouns.

Bardina (2003) interpretes adjectivization as the result of a two-stage interaction that begins with semantic and ends with syntactic derivation. Semantic derivation gives rise to new semantics, and consists of splitting a word form into two homonyms within one lexical grammatical class (ibid.: 72). Syntactic derivation produces new properties in a grammatical category, as it shifts from one grammatical class to another. The process of the transition from one POS to another begins when a lexeme is used as another POS in atypical contexts. In the course of this process, a participle is used as a homonym of an adjective. Thus, novel (that is, adjectival) uses of participles result from the semantic shift, and reflect the workings of syntactic change. The transition results in the formation of a new lexeme with a different categorical meaning from the original lexeme. Adjectivization results from this interaction after semantic change occurs, including an obligatory change in the denominative meaning of a participle.

The main differences between the syntactic and morphosemantic approaches lie in the nature of the processes. The syntactic approach, which focuses on adjectivization as a side-by-side process of semantic change and the loss of verbal properties, does not explain the causes of these changes, and attributes the factors to adjectivization only given the properties of the syntactic context. The morphosemantic approach relies on two steps, which are the gradual extension of lexical semantics (realized by the prerequisites) and a change of grammatical meaning that takes place first, followed by the qualitative grammatical shift leading to homonymy. In this approach, the syntactic context highlights the result of adjectivization rather than its properties.

4.2.1.1 Semantic relationship between verbs and adjectives

Kustova (2012) maintains that adjectivized participles are semantically related both to adjectives via their synonymous adjectival meaning and to their base verbs via the semantic property of verbal event. Verbal event is a type of event denoted by the base verb, and conveys

[50] It is the process whereby a noun, an adjective, a numeral, a gerund, and so on becomes an adverb; for example, *šepotom* ‘in a whisper’ (N ⇒ ADV), *veselo* ‘joyfully’ (ADJ ⇒ ADV), *vdvoem* ‘together’ (NUM ⇒ ADV), and *nexotja* ‘reluctantly’ (adverbial participle ⇒ ADV; Babajceva 1967).

an action performed once or repeated several times, or a state of being repeated (several times) or performed once. Verbal event can also apply to states conveyed by adjectivized participles; for example, *brošennye doma* 'abandoned homes' (from *brosit′* 'abandon'), and *obvetrennoe lico* 'weather-beaten face' (from *obvetrit′* 'make weather-beaten'). Kustova (2012) does not state explicitly whether a verbal event applies to both action and state verbs.

In this study, the endpoint of adjectivization is when an adjectivized lexeme no longer functions as a verbal form but remains connected to a verbal event. The complete loss of this connection, as in *blestjaščij muzykant* 'brilliant musician' (an adjectivized participle), is viewed as an extreme of adjectivization; it is also manifested in metaphorical and metonymical[51] uses (e.g., Černega, 2009). The loss can be a formal distinction between an adjectivized participle that maintains or lacks relatedness with to the verbal meaning (see Section 4.1.4). For example, there is no direct relationship between the meaning of being brilliant, as in *blestjaščij muzykant*, and the meaning of reflecting light, as in the verb *blestet′* 'shine'. By contrast, in *blestjaščie volosy* 'shiny hair', there is still a link between the meaning of the base verb *blestet′* with the meaning 'to shine' and its adjectivized participial word form *blestjaščie* 'shiny, be able to shine in the light'.

A verbal event affects the meaning of an adjectivized participle according to which they can be classified as the two most numerous groups, namely active present adjectivized participles (or ADJP$_{\text{ACT}}$, such as *rukovodjaščij rabotnik* 'leading employee') and passive past participles (or ADJP$_{\text{PASS}}$, as in, *predubeždennyj vzgljad* 'biased outlook'). ADJP$_{\text{ACT}}$ and ADJP$_{\text{PASS}}$ convey new meanings that language users cannot otherwise communicate using existing means of communication. This implies that present active and past passive participles may become adjectivized more often because they fill gaps in the lexicon, in addition to regular adjectives.[52] For example, the present active participle *rukovodjaščij* is used as an adjective *rukovodjaščij rabotnik* 'leading employee', conveying an additional meaning of someone who manages or is in charge of something, and applying to the narrow context of the work environment.[53]

Most ADJP$_{\text{ACT}}$ are relational adjectives[54] that may have an agentive component; for example, in *uvlavžnjajuščij krem* 'moisturizing cream' – a cream that moisturizes skin – the first participant is cream and the second participant is the skin affected by the cream. The meanings of ADJP$_{\text{ACT}}$ are mainly related to a verbal event that is repeated multiple times, as in *igrajuščij trener* 'playing coach', *kormjaščaja mat′* 'feeding mother', and *p′juščij sosed* 'alcoholic neighbor'.

An ADJPs$_{\text{PASS}}$ (with a relational and qualitative meaning) conveys a real or an imaginary verbal action that could only be expressed otherwise by the semantics of synonymous adjectives

[51] I may refer further to *idiomatization* and *idiomatized* meanings to convey both the metaphorical and the metonymical uses.

[52] Kustova (2012) does not state explicitly whether these adjectivized participles compete with adjectives or whether adjectivization occurs due to language users' needs to convey new meanings.

[53] Some of the few synonymous adjectives (compared to the larger number of their corresponding participles) for *rukovodjaščij* are *glavnyi*'principle, chief' and *central′nyj* 'central'. Either of these adjectives used with the noun *rabotnik* 'employee' characterize the employee as important or as being the main focus, but do not specify that he is important with regard to the tasks in which he is in charge.

[54] Kustova (2012) notes that there are fewer qualitative adjectives of this type.

(such as *brošennye doma* 'abandoned homes' – *pustye doma* 'empty homes').[55] An ADJP$_{\mathrm{PASS}}$ references verbal events in the past, mainly those that occurred once, continuously, or repeatedly (optional); for example, *iznošennoe pal'to* 'worn coat', referring to a coat that has been worn for a long time. Referencing the recurrence of verbal event reflects the perfectivity and resultativity of adjectivized participles and their base verbs.

An ADJP$_{\mathrm{PASS}}$ conveys lexical meanings that fewer differences in contrast to the meanings of their respective synonymous adjectives. The generalized lexical meanings are grouped according to five models, namely 'motivating event', 'perfect', 'comparative model', 'quantitative model', and 'qualitative model'. All the models convey different aspects of meanings in adjectivized participles that direct them towards becoming clear-cut adjectives. The adjectivized participles in these models can specify general notions (motivating event), share an additional meaning of completion (perfect model), or shift from one property to another (comparative model). The meanings of the participles can also be extended to qualitative meanings conveying quantities (quantitative model) and additional negative, positive, and neutral connotations (qualitative model).

In the model of motivating event, the meaning of ADJP$_{\mathrm{PASS}}$ highlights a subclass of the general class; for example, a *priglašennyj professor* 'invited professor' is one type of professor, who may also receive a salary, not be a member of the permanent staff, and so on. In the perfect model, ADJP$_{\mathrm{PASS}}$ conveys the perfective meaning and references a significant state arising from the preceding event (perfective meaning); for example, *okrašennye volosy* 'dyed-PP.PFV hair'[56] derived from *krasit* 'dye', synonymous with *krašenye volosy* 'dyed-ADJ hair', which refers to hair that is not its natural color.

The comparative model, which is the most numerous of the models, includes ADJP$_{\mathrm{PASS}}$ that conveys a small shift in the attributive/scale meaning related to physical objects, such as parts of the body, plants, and the like. For example, the adjectivized participle *uproščennaja* in *uproščennaja procedura* 'simplified procedure' conveys the meaning of 'relatively simple, easier than it could have been', which indicates that there is a shift on the scale of property from 'simple-difficult'. The participle is synonymous with the adjective *prostoj* 'simple', which does not convey a shift on the scale from simple to difficult. The quantitative model includes adjectives containing the semantic component 'a lot of, large number'. For example, an adjectivized passive participle with a relational meaning can be extended to the qualitative meaning of an adjective. The word form *osvedomlennyj*, as in *osvedomlennyj istočnik* 'informed source', conveys a source with a large amount of information; therefore, this meaning suggests that *osvedomlennyj* is closer to being a qualitative adjective.

Lastly, in the qualitative model, the ADJP$_{\mathrm{PASS}}$ word form *osnaščennaja* 'equipped', as in

[55] ADJP$_{\mathrm{PASS}}$ aare not always related to real passive participles formed by transitive verbs, and many of them can be related to reflexive verbs; for example, *učaščennyj* 'quick', *iskrivlennyj* 'crooked', and *vytjanutyj* 'oblong, stretched'.

[56] The aspectual meaning of *okrašennye* characterizes the quality of the hair (such as dry, fragile, and so on) as a result of dying it.

osnaščennaja labaratorija 'equipped laboratory', characterizes both the quality of the equipment (the laboratory has the necessary, functioning equipment) and the availability of favorable conditions for successful research (or other purposes, depending on what a particular laboratory is used for). This is why *osnaščennaja* conveys a judgmental meaning based on implications about quality from the verbal action, and highlights positive characteristics. The meaning of this type of ADJP$_{\text{PASS}}$ is focused more on the (negative/neutral/positive) characteristics themselves than it is on judgments (Kustova, 2012).

Kustova (2012) notes that the most prevalent types (mentioned above) are the comparative, quantitative, and qualitative models because they contain semantic properties that are not available in other classes of adjectives. Thus, they satisfy the communicative needs of language users. Kustova's (2012) approach demonstrates how adjectivized participles are connected to regular adjectives by means of semantics, and how they offer more possibilities for using different semantic properties to fill gaps in communication. This correlation of the semantics of adjectivized participles and adjectives excludes syntactic factors, and does not explain why, apart from the pragmatic need for another means of communicating a new meaning, the lexeme becomes adjectivized.

4.2.2 Causes and stages of adjectivization

This section focuses on the typology of participles based on their grammatical categories and the semantics of their base verbs. This typology enables the definition of the word-internal properties that account for adjectivization and identify its stages; in other words, what triggers adjectivization and how it manifests formally in a syntactic context. The approach based on the word-internal properties that lead to adjectivization was highlighted by Kolochkova (2011), Kalakuckaja (1971), Kustova (2012) and Černega (2009). The approach differs from Say's (2016) perspective on adjectivization, which only considers the loss of verbal properties and the shift in semantics as processes underlying the factors of adjectivization. The actual causes of adjectivization remain undefined in the syntactic approach. Kolochkova, Kalakuckaja, Kustova and Černega argue that morphosyntactic properties (such as tense, aspect, voice, and transitivity) and the semantic properties associated with them (such as processuality and resultativity), figurative meanings and the meaning of concrete action are inherited from a verb and can either favor or disfavor adjectivization. Kolochkova (2011) holds that adjectivization depends on the degree of verbality, which can facilitate or complicate the transition into the class of adjectives.

I reviewed approaches discussed by Kalakuckaja (1971) and Černega (2009) to answer the following questions: To which morphosyntactic and semantic properties of participles and their base verbs can adjectivization be attributed? Where do these properties originate? What is the role of a base verb in the development of adjectivized properties within a participle? Exploring these questions may also reveal whether the internal or contextual properties of a participle affect its predisposition to and degree of adjectivization.

Kalakuckaja's (1971) study provides a detailed account of the factors that create a predisposition towards adjectivization, restrict it, or do not prevent it. These factors represent the internal morphological and syntactic properties of participles together with the semantics of their corresponding base verbs. Kalakuckaja assumes that the potential for adjectivization is attributed to the morphosemantic factors, while the effect of the syntactic context is secondary and is only exerted on 'individual adjectivization' (ibid.: 74). This implies that adjectivization begins with a change in meaning induced by the grammatical and semantic properties of participles, and is manifested in the syntactic context. On a more general level, it implies that adjectivization is a process in the course of which verb semantics and the meanings of grammatical categories determine syntactic behavior. One should note that Kalakuckaja (1971) discusses these properties based on her intuition and information from grammars of Russian, and supports her claims by citing examples from literary texts. Although no corpus evidence was used to test these claims, her approach, which relies heavily on evidence from Russian grammars and a thorough analysis of examples, appears to be well argued. In addition, her approach may explain the initial cause of adjectivization discussed in the syntactic approach (Section 4.1).

Table 3.2 provides a brief summary of the morphosyntactic and semantic properties that favor (+ adjectivization) or disfavor (- adjectivization) adjectivization. These properties are attributed to four types of participles, namely present active participles (ACT PRS PTCP), past active participles (ACT PST PTCP), present passive participles (PASS PRS PTCP), and past passive participles (PASS PST PTCP).

Properties	**+ adjectivization**	**- adjectivization**
ACT PRS PTCP	• the first most numerous group of adjectivized participles[57] • unmarked/unrestricted meaning of **present** tense: temporal meaning transforming into atemporal • figurative meaning of base verb • **intransitive** verbs have less possibilities for verb government	• meaning of concrete action conveyed by base verbs • strong predisposition of derived **transitive** verbs for verb government • the availability of the reflexive affix *-sja-* (for **intransitive** verbs) • formal markers of imperfective aspect (suffixes or prefixes) • **active** voice • unmarked **imperfective** aspect: processual meaning, close connection to the verbal system

Table 3.2 – continued on the next page

[57] This applies to active present imperfective participles.

ACT PST PTCP [58]	• **perfective** aspect, **past** tense: resultative meaning of action • productive suffixes -*š*- conveying qualitative meaning • **perfective intransitive** base verbs without ungrammaticalized suffixes and the suffix -*nu*- • no effect from lexical prefixes; for example, *zanošennyj pidžak* 'worn down suit' (prefix *za*-)	• **imperfective** aspect:[59] processual temporal meaning, close connection to the verbal system • **past** tense, **imperfective** aspect: processual meaning • restriction and markedness of **past** tense is stronger than is that of **present** tense: temporal meaning • lexical affixes conveying meaning of concrete action • **active** voice • **perfective** suffix -*nu*-, for **active past** participles
PASS PRS PTCP	• marked **passive** voice: no government of **accusative** nouns without prepositions [60] • **perfective** aspect • base verbs conveying the property of an object ⇒ possibility of being used with adverbs of measure and degree; for example, *samye uvažaemye graždane* 'the most respected citizens'	• productive base **imperfective transitive** verbs • meaning of concrete action conveyed by base verbs; for example, *reguliruemaja lampa* 'adjustable lamp', *otvinčivaemaja gajka* 'a screw-nut (that can be unscrewed)' • secondary and unproductive **imperfective** base verbs; for example, *zatračivaemaja ènergija* 'spent energy', *privivaemyj bol′noj* 'vaccinated patient' • use in syntactic contexts (with **instrumental** nouns) • stylistically limited use (business, scientific style), bookish style with the participial suffix *-(m)yj-* • negation particle *ne* 'not'

Table 3.2 – continued on the next page

[58] Kalakuckaja (1971) argues this type of participle is the least inclined towards adjectivization.

[59] Kalakuckaja (1971)) argues that the adjectivization of past active participles derived from imperfective verbs is impossible, apart from one exception: *byvšij* 'former', synonymous with the adjectives *prošlyj* 'last' and *staryj* 'old'.

[60] The government of accusative nouns is a typical verbal property, and its absence is a sign of no correlation between the voice of participles and the voice of finite verbs.

PASS PST PTCP[61]	• the second most numerous group of adjectivized participles • **perfective** aspect (having no correlation with **imperfective** forms) and **passive** voice ⇒ strong resultative meaning of **past** tense • **passive** voice does not prevent adjectivization • figurative meaning of **perfective** base verb; for example, *ubityj vzljad* 'extremely tired look' • resemblance of participial suffixes *-nn-* and *-t-* with adjectival suffixes *-n-, -at-, -ovat-* • rarely used with agents	• meaning of concrete action of the base verbs (decreases the qualitative meaning); for example, *napisannyj* 'written', *sšityj* 'sewn' + lexical affixes[62], such as *pridvinutyj stol* 'a table pulled closer (to a wall or door, for example)' • restriction and markedness of **past** tense is stronger than **present** tense: temporal meaning

Table 3.2: Classification of participles with regard to their semantic and morphosyntactic properties that favor or disfavor adjectivization, based on Kalakuckaja's (1971) and Kustova's (2012) classifications, with some additional information from Černega (2009).

Table 3.2 shows that the present tense, the passive voice, the perfective aspect, and the intransitive use can distance a participial word form from the verbal paradigm and move it closer to the adjectival one, mainly through the extension of the semantics. Intransitive participles are more inclined towards adjectivization due to their reduced argument structure, which requires one subject argument, compared to a transitive structure that requires both a subject and an object (e.g., Allen, 2009).

Although the passive voice is a strongly marked[63] category, it is not used in predicative finite verbal forms; hence, the passive participles are more distant from predicative verbal forms than active participles are. In addition, passive participles can only govern a noun/pronoun in the instrumental case without a preposition. These two properties, conveyed in the grammatical meaning of passive participles, tend to move them away from the verbal paradigm.

The perfective aspect conveys the result of action qualifying this action. Past passive perfective participles convey a highly resultative meaning, as their corresponding base verbs are not used as passives; see the examples in (29) by Kalakuckaja (1971: 177). The active participial word form *razbivšij* 'having broken' in (29b) has a corresponding active verbal word form *razbil* 'broke' in (29a), while the participial word form *razbityj* 'broken' in (29c) has no corresponding finite

[61] They are heavily inclined towards adjectivization.

[62] They specify the degree and quality of the process of action.

[63] *Markedness* "presupposes the notion of formal complexity, whereby the marked is structurally more complex and the unmarked more simple" (Givón, 1991: 3). The marked category is usually less frequent (and thus cognitively more salient), and the unmarked is more frequent (ibid.).

verbal form in the passive voice.

(29) a. *razbil*
break.PST
'[he] broke'

b. *razbivšij*
have.break:PP.ACT
'[he] having broken'

c. *razbityj*
break:PP.PASS
'[it is] broken'

As opposed to adjectivized participles that are part of the adjectival paradigm, unambiguous participles maintain the verbal property of transitivity.[64] Transitivity accounts for the formation of morphological types of participles and verbal government. First, transitive verbs form active and passive particles, while intransitive verbs only form active participles. Second, intransitive forms have a reduced argument structure (one argument) which, according to this criterion,[65] generally prevents them from taking an object. For this reason, in addition to the effect of voice on verb government discussed by Kalakuckaja (1971), the reduced argument structure in **active intransitive participles** (without strong verb government) leading to a lack of complements may be another factor favoring adjectivization.

The past tense, the imperfective aspect, the active voice, and transitivity obstruct adjectivization by maintaining connection with the verbal paradigm. The past tense is more strongly marked and, unlike the present tense, remains within the verbal paradigm without shifting towards the adjectival one. The active voice is more strongly marked because it is firmly associated with finite verbal forms, presupposing the government of nouns in the accusative case. Transitive verbs are strongly predisposed for verb government, and have a more complex argument structure (more than one argument). Syntactically, this manifests in the availability of complements in the immediate context.

The initial semantics of base verbs also affect the predisposition towards adjectivization: Verbs with figurative, abstract or qualitative meanings allow a participial word form to become more adjectival, while verbs conveying concrete action obstruct adjectivization. The examples in (30) illustrate the participial and adjectival homonyms *ottalkivajuščaja* 'bouncing off/repugnant'. The difference between the word forms is based on the meanings of the base verb *ottalkivat′* 'push off, repulse'. The literal one denotes a concrete action of bouncing/pushing off something, as shown in (30a) in which *ottalkivajuščaja* 'bouncing off' is a participle used with the direct

[64] Transitivity implies transitive and intransitive uses of participial forms and other verbal forms (that is, finite forms, infinitives, and adverbial participles). I will further refer to 'the transitive use' and 'the intransitive use' as specific properties of a verbal form relating to transitivity.

[65] This statement does not imply that all intransitive verbs will not necessarily select a subject solely because of their one-argument structure. Some intransitive verbs are still capable of strong verb government and require complements in the oblique cases, with or without prepositions (for example, *èto mešalo mne* 'it bothered me', in which the word form *mešalo* is derived from the intransitive verb *mešat′* 'bother'; Bořkovec 1976).

object *mjači*k 'ball'. The qualitative meaning denotes the state of being unpleasant and repugnant. The example in (30b) illustrates that the word form *ottalkivajuščaja* 'repugnant' is used as an adjective that modifies *vnešnost′* 'appearance' as being repugnant.

(30) a. *sila, ottalkivajuščaja mjačik ot stenki*
force bounce:PRESP ball off wall
'the force, bouncing the ball off the wall' (RNC)

b. *ottalkivajuščaja vnešnost′*
repugnant.NOM.SG.F appearance
'repugnant appearance'

The examples in (31) show participial and adjectival word forms derived from the base verb *padat′* 'fall' with a concrete meaning of falling and an abstract meaning of sinking, depressive or being depressed (for example, when referring to a psychological state). In (31a), the participial word form *padajuščie* 'falling' is used with a concrete noun *zvezdy* 'stars', while in (31b), *padajuščij* 'heavy' is an adjective that modifies the abstract noun *dux* 'heart', conveying a psychological state.

(31) a. *padajuščie zvezdy*
fall:PRESP stars
'falling stars'

b. *padajuščij dux*
heavy.NOM.SG.M spirit
heavy heart

Kalakuckaja (1971) argues that certain types of participles are inclined to be adjectivized depending on their affixes. For example, prefixes in present participles often specify an action, which in turn requires explanatory words (for example, spatial modifiers), which then obstruct adjectivization. In (32), the present participle *vosxodjaščim* 'raising' with the imperfective prefix *vos-* is used with the adverb of manner *tixo* 'quietly' and the prepositional phrase *nad . . . rekoj* 'over . . . the river' as the adjunct. The presence of the adverb and the adjunct in the syntactic context is conditioned by the imperfective prefix *vos-* in this participial word form.

(32) *s ego vosxodjaščim tixo dymom nad vysyxajuščej rekoj*
with its raise:PRESP quietly smoke over diminishing river
'with its quietly rising smoke over the diminishing river' (Kalakuckaja, 1971: 82)

Grammatical prefixes in past passive participles do not obstruct adjectivization, while the suffixes *-nn-* and *-t-*, which resemble the adjectival suffixes *-n-, -at-*, and *-ovat-*, opt for it. For example, the word form *otstranennyj* with the suffix *-nn-* can be a participle denoting 'suspended (from duties)' or 'aloof', and the word form *razbityj* with the suffix *-t-* can be a participle with the meaning of 'broken' or an adjective meaning 'frustrated'. Affixes that disfavor adjectivization are the reflexive affix *-sja-* in the present active intransitive forms of participles (for example, *izmenjajuščijsja* 'changing') and the perfective suffix *-nu-* in past active participles (such as, *pokinuvšij* 'having left'). The lexical meaning of the suffix *-nu-* specifying a one-off action prevents past active perfective participles from adjectivization (Kalakuckaja, 1971: 133).[66] In addition, the negation particle *ne* 'not' is said to resist adjectivization. I consider these morphological properties to be connected indirectly to adjectivization because the primary assumption is that semantics and the grammatical meanings of participles undergo changes in the course of adjectivization. Table 3.3 below provides a summary of morphosemantic properties favoring and disfavoring adjectivization derived from Table 3.2.

Properties	**+ adjectivization**	**- adjectivization**
Description	Semantics of base verbs and internal morphosyntactic properties of participial word forms favor adjectivization. They decrease the connection with the verbal paradigm and draw it closer to the adjectival one.	Semantics of base verbs and morphosyntactic properties of participial word forms maintain a connection with the verbal paradigm and prevent the word forms from acquiring adjectival properties.
Semantics of the base verbal lemma	• figurative meaning • abstract meaning • meaning denoting quality or property (combined with an attributive function)	• meaning of concrete action decreasing qualitative meaning (typical of adjectives)
Morphosyntactic categories of participles	• **passive** voice excluding the government of **accusative** nouns without prepositions • **present** tense conveying an atemporal meaning through the simultaneity of action • **perfective** aspect conveying resultative meaning • **intransitive** verbs with fewer possibilities of verb government	• **active** voice • **past** tense conveying a more defined temporal meaning due to its markedness • **imperfective** aspect conveying processual meaning including repetition and recurrent actions (see Černega 2009) • **transitive** verbs strongly predisposed for verb government

Table 3.3 – continued on the next page

[66] Past active perfective participles without the suffix *-nu-* or lexical prefixes can be adjectivized; for example, *minuvšij* 'last', as in *minuvšej zimoj* 'last winter'.

Morphological properties of participles	• **past active** participles: productive suffixes *-vš-*, *-š-* • **present active** participles: **perfective intransitive** base verbs without lexical prefixes or the suffix *-nu-* • **past passive** participles: resemblance of participial suffixes with adjectival ones	• the reflexive affix *-sja-* in **present active** participles has a strong relation to the subject and object of the action • more defined temporal meaning of the **past tense** due to its markedness • processual meaning of the **imperfective** aspect • **present passive** participles: stylistically limited use (business, scientific style), bookish style with the participial suffix *-(m)yj-*, negation particle *ne* • **past active** participles: **perfective** suffix *-nu-*, lexical suffixes

Table 3.3: Classification of semantic and morphosyntactic properties favoring and disfavoring adjectivization.

These properties are internal, and affect adjectivization to the point at which an adjectivized participial word form behaves syntactically as an adjective. In this regard, the syntactic context only reflects whether a participle is adjectivized under the influence of its morphosemantic properties, which favor or do not favor adjectivization. This is in line with Kalakuckaja's (1971) claim that adjectivization is a grammatical property of participles and does not depend on the syntactic properties of a sentence in which a participle is used. According to this view, the syntactic context may intensify the adjectivization of a participle triggered by some of the internal properties favoring adjectivization, as shown in Table 3.3. It is still unclear whether the syntactic context can further influence the course of adjectivization once a participial word form has been used as an adjective. At a more general level, the outcome of adjectivization is a lexeme that has a qualitative meaning and which behaves syntactically as an adjective.

If we assume that adjectivization arises from the internal properties of the base verbs and participles, and becomes realized in a syntactic context, there should be an intermediate level at which a participle undergoes some changes prior to behaving syntactically as an adjective. To investigate this, I devised a scheme of adjectivization that included several levels and stages (see Figure 3.1). The scheme illustrates the connection between the semantic and grammatical properties of participles and verbs, their adjectivization, and subsequent use in the syntactic context (that is, their syntactic behavior). I consider a word's internal properties to be the basis for adjectivization, leading to changes within participial semantics, argument structure, and the ability of verbal government. For example, a participle that is inclined towards adjectivization can inherit a figurative meaning from its base verb, and can be in the present tense and the passive voice. It is not certain whether a participle with a figurative (or any other similar lexical) meaning

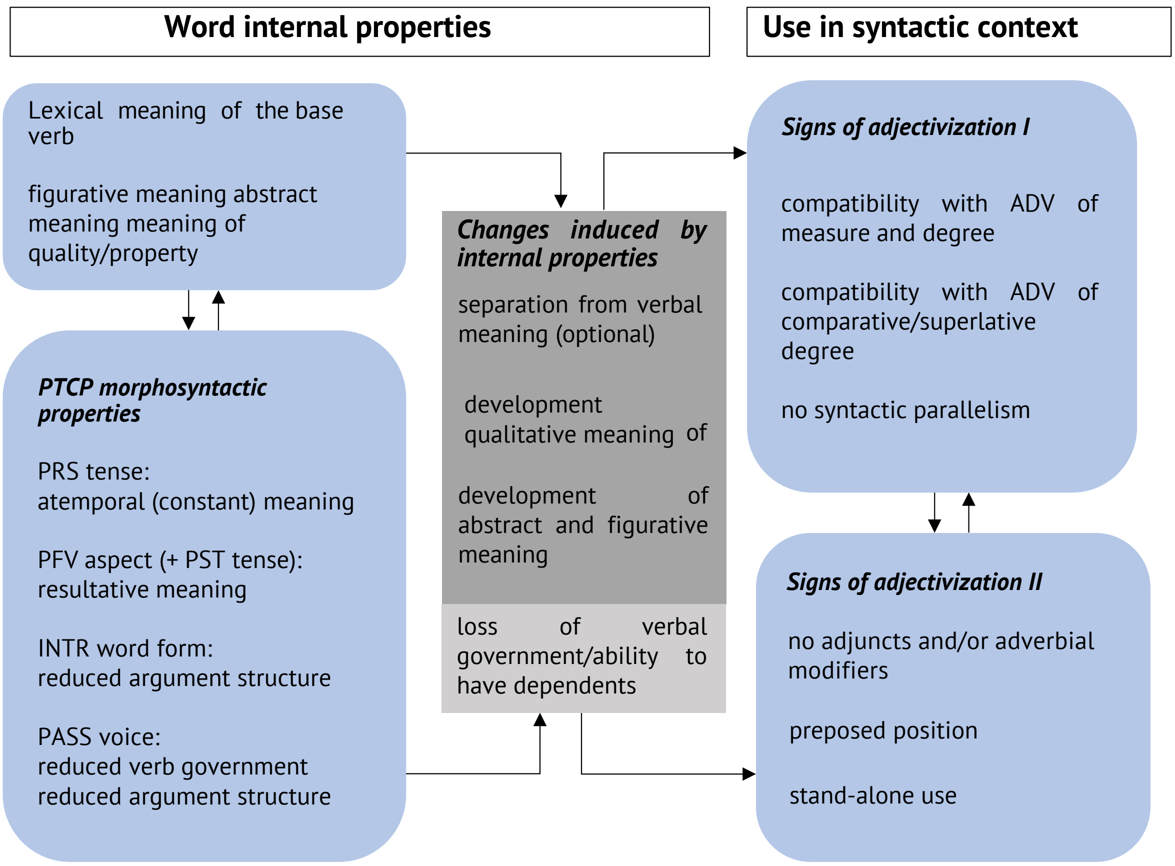

Word internal properties
Use in syntactic context
Lexical meaning of the base verb
figurative meaning abstract meaning meaning of quality/property
PTCP morphosyntactic properties
PRS tense: atemporal (constant) meaning
PFV aspect (+ PST tense): resultative meaning
INTR word form: reduced argument structure
PASS voice: reduced verb government reduced argument structure
Changes induced by internal properties
separation from verbal meaning (optional)
development qualitative meaning of
development of abstract and figurative meaning
loss of verbal government/ability to have dependents
Signs of adjectivization I
compatibility with ADV of measure and degree
compatibility with ADV of comparative/superlative degree
no syntactic parallelism
Signs of adjectivization II
no adjuncts and/or adverbial modifiers
preposed position
stand-alone use

Figure 3.1: Scheme of adjectivization.

also tends to lose the ability to join verb complements. I suggest that the figurative meaning should enable the qualitative meaning, in order for a participle to be used as an adjective in the syntactic context. This use may prevent it from having dependents or governing nouns because an adjective does not have these properties.

The changes induced by **internal properties** represent the first stage of adjectivization. At this stage, the meaning of a participle diverges from the lexical meaning of the base verb and becomes qualitative or abstract/figurative. It can also become idiomatized, thus separating it from the verbal meaning. The participle also loses the ability to have verb dependents and verb government (influenced by the morphosyntactic properties). The following stage of adjectivization reflects how an adjectivized participle comes to behave in a **syntactic context**, and is based on the signs of adjectivization discussed in the syntactic approach. *Signs of adjectivization (I)* are assumed to result from changes in the lexical meaning. For example, the lexical meaning of the participle becomes qualitative, and can therefore combine with adverbs and adjectives of comparative/superlative degrees. *Signs (II)* result from the loss of verbal government and a reduction of the argument structure. For example, the atemporal and resultative meaning of the present tense and the perfective aspect can resist spatial/temporal modification, so that the adjectivized participle will not be combined with adjuncts. The intransitive use and the passive voice prevent it from having verb dependents and certain types of verb government; the intransitive and/or passive participles will therefore be used without syntactic markers of adjectivization.

4.3 Summary

In this section, I discussed morphosemantic approaches to adjectivization based on studies such as those by Kolochkova (2011); Kalakuckaja (1971); Kustova (2012) and Černega (2009). These focus on the change in semantics that accompanies changes in the grammatical meanings of tense, aspect, transitivity, and voice. On one hand, these changes result in POS homonymy; on the other, they are reflected in the adjectival syntactic behavior discussed in the syntactic approach. These changes arise from both the lexical semantics inherited from the base verb, and from the grammatically induced meanings in tense, aspect, voice, and transitivity.

The most common type of adjectivized participles are active present and past passive participles, both of which allow language users to convey meanings that cannot be conveyed by synonymous adjectives (Kustova, 2012). This may potentially explain why native speakers have come to use participles as adjectives: These participles convey new meanings that reflect a connection with the verb event and adjectival relational/qualitative meanings. This also reflects the internal grammatical tendency of participles to become adjectivized despite the context in which they are used (e.g., Kalakuckaja, 1971).

Internal properties of participles and their verbs affect their ability to become adjectivized: some of these properties favor it, while others do not. These properties concern the lexical semantics of verbal lemmas, the grammatical categories of tense, voice, aspect, the transitivity of participles, and particular affixes and suffixes associated with certain types of participles (see Table 3.2). First, the grammatical meanings of the present tense and the perfective aspect (also combined with the past tense) enable the development of a qualitative meaning and detachment from a verbal meaning. The present tense conveys an atemporal meaning, which can dissociate from temporality and come to denote a state rather than a process. The perfective aspect denotes a result characterizing an action, which brings it closer to having the meaning of an adjective. Second, the passive voice and the intransitive use presuppose the reduced argument structure of a participle and a limited ability to join complements. Syntactically, this is manifest in an absence of adjuncts and complements. Third, the lexical meanings of base verbs (figurative, abstract, and qualitative) lead to their extension in participles, which draws the semantics of a participle closer to those of an adjective and provokes adjectivization.

Although the internal properties of a participle affect its predisposition to adjectivization, the syntactic context in which a participle is used reflects the effect of these properties. Changes in the semantics of the participle and the effect of its morphosyntactic properties (tense, aspect, voice, and so on) lead to the syntactic behavior of the adjectivized participle being manifested in its compatibility with adverbs of measure/degree, stand-alone use, and the absence of adverbial modifiers, for example (Figure 3.1). The properties that disfavor adjectivization allow a participle to combine with complements and adjuncts, which is a direct sign of adjectivization. On the other hand, the properties favoring adjectivization lead to the stand-alone use of a participle (for example, present passive participles that lack an instrumental noun).

Thus, a participle can detach from the verbal paradigm and become an adjective if its base verb already has an abstract, figurative, or qualitative meaning, and/or it has a set of morphosyntactic properties (such as the present tense or the passive voice). These properties facilitate semantic change (the development of adjectival meaning) and the reduction of arguments and verb government; these changes result in the adjectivized lexeme being used as an adjective rather than as a participle in syntactic contexts.

5 Conclusion

In this chapter, I investigated the properties of POS ambiguity, type 2 (honomyms with related morphological forms and meanings) from the synchronic perspective. This type of ambiguity is less concerned with morphological processes compared to the intra- and inter-paradigmatic ambiguity. It is a specific type of conversion, and is a universal process of derivation that triggers a change in syntactic function (as the first stage) that results in a change in semantic properties without bringing about changes in the paradigm, or graphical and phonological forms of a lexeme (e.g., Manova, 2011). This process is an affixless type of word formation that is typical of isolating and analytic languages such as Chinese and English, and atypical for Russian, which prefers affixation as the most productive means of word formation.

Adjectivization was approached from the syntactic (Say, 2016; Timberlake, 2004) and morphosemantic perspectives (Kustova, 2012; Kalakuckaja, 1971; Kolochkova, 2011; Černega, 2009). These perspectives differ in terms of the syntactic context of an adjectivized lexeme versus its semantic/morphological properties. The syntactic approach focuses on the properties of the immediate syntactic context surrounding the adjectivized lexeme, and explains how these properties allow for the distinction between adjectivized and unambiguous participles. Therefore, this approach explains how adjectivized and unambiguous participles can differ based on the syntactic context in which they are used. The semantic approach relies on the grammatical and lexical meanings of a participle with regard to its ability to favor or resist adjectivization. The central role is attributed to the change in the grammatical and semantic meanings of a participle that trigger or obstruct the subsequent development of adjectival properties and the loss of verbal properties. Therefore, certain morphosemantic and syntactic properties of participles that cause or resist adjectivization are primary in this approach; the syntactic context only illustrates the use of adjectivized or unambiguous participles as an effect of these properties. The end result of adjectivization is viewed similarly in these approaches: In the syntactic approach, a participle is fully adjectivized when its semantics are no longer related to the semantics of its base verb. In the morphosemantic approach, adjectivization results in homonymy between the adjectivized participles and their corresponding adjectives.

Adjectivization can be viewed as a process wherein participles lose their syntactic/semantic properties shared with finite verbal forms and become part of the class of adjectives (Say, 2016). This process relies on one or more syntactic and semantic factors that affect and/or reflect the degree of adjectivization. These factors can be both morphosyntactic and semantic properties of participles, and their syntactic behavior manifests itself through these properties. Unlike unambiguous participles, adjectivized participles have reduced argument structure, loose connection with the semantics of base verbs, and a lack of spatial/temporal modification. In addition, they also cannot be paraphrased as relative or finite clauses. Furthermore, attributive and predicative functions do not play prominent roles in differentiating between verbal and adjectival uses of participles. One exception is short-form present passive participles used predicatively that do not adjectivize (for example, *on ljubim* 'he is appreciated-PP') and full-form present active participles used predicatively as arguments of semi-linking verbs that do adjectivize (for example, *rezul'taty okazalis' ošelomljajuščimi* 'the results turned out to be amazing-ADJ' (RNC)). The properties that seem relevant for adjectivized rather than unambiguous participles are a preposed (and attributive) position, use with adverbs of measure/degree, with adverbs of comparative/superlative degree or superlative adjectives, and stand-alone use (no arguments or adjuncts). The overlap of these properties does not signify that they are no longer factors of adjectivization because the syntactic context should meet at least one of these criteria for the word to be adjectivized. The meaning of each adjectivized participle undergoes a lexical shift, to a greater or lesser extent.

I classified participial lexemes as unambiguous participles, unambiguous deverbal adjectives, and adjectivized participles that are part of the verbal paradigm or are homonymous with the verbal forms. Such a classification draws a clear line between participial lexemes affected/not affected by adjectivization, which is useful for defining the profile of participles that require disambiguation using the CG formalism.

The morphosemantic approach focuses on the internal properties of participial lexemes that favor or disfavor adjectivization. Therefore, the essential property of adjectivization is the modification of semantics (extension or specification; Bardina 2003), the loss of verbal meaning (the temporal property of action), and the reinforcement of the qualitative meaning intrinsic to adjectives. The morphosemantic properties of participles may affect their predisposition to adjectivization and their syntactic behavior towards that of an adjective.

Active present and past passive participles represent the largest group of adjectivized participles (Kustova, 2012) disposing of internal morphological and semantic properties that generally favor adjectivization. Tendencies towards adjectivization are strong in present active participles, whereby the meaning of the present tense can transform into atemporal adjectival ones; the intransitive use implies a lack of verb government and does not allow for the selection of objects. Past passive participles derived from perfective verbs are most likely to acquire a qualitative meaning because the passive voice and the past tense tend to lose temporal meaning and processuality due to the resultative meaning of the past tense and specific suffixes that are homonymous with adjectival suffixes; for example, the suffixes *-nn-* and *-t-*. The group of present active and past passive adjectivized participles supplies the class of adjectives with new means of expressing lexical meaning that (a) existing adjectives cannot express and (b) are synonymous with the meanings of existing adjectives. This is the most numerous group of adjectivized participles because it fills gaps in communication by offering new meanings that are unavailable in the existing class of adjectives.

Chapter 4

Exploratory analysis

1 Introduction

In this chapter, I conduct a quantitative analysis of the factors of adjectivization discussed in Chapter 3. In the analysis, I consider (1) syntactic, (2) semantic, and (3) morphosyntactic factors, as shown in Table 4.1. I also introduce the additional factor of frequency (4) that has not been discussed previously with regard to adjectivization in the reviewed literature.

#	Factors	Description
1	**syntactic context**	compatibility of a participle with an adverb of measure and degree: *očen′* ‘very’
2	**semantics**	semantic classes of the base verbs and their corresponding participles
3	**morphosyntactic properties**	tense, voice, aspect, and the transitivity of a participle
4	**frequency**	corpus frequency of base verbs and their corresponding participles

Table 4.1: Factors of adjectivization investigated in the exploratory analysis.

The compatibility with *očen′* ‘very’ (factor 1) is studied across two types of adverbial constructions, *očen′* ‘very’ used with participles (for example, *ochen′ podhodjashaja* ‘very becoming’) and *očen′* ‘very’ used with their corresponding finite verbs (for example, *ochen′ podhodit* ‘[. . .] fits a lot’). The comparison of the constructions is based on the corpus frequencies of participles and verbs. Semantics (factor 2) is represented by the semantic classes of base verbs that are inherited by their corresponding participles. For example, the participle *podhodjashaja* ‘suiting’ maintains a semantic affinity with the base verb *podhodit′* ‘fit, suit’ as a verb of mental domain. The morphosyntactic properties of a participle (factor 3) may favor or disfavor adjectivization, as discussed in the morphosemantic approach. Corpus frequency (factor 4) is represented by the corpus frequency of the base verbal lemmas and the ratio of participles to finite/infinitival verbal word forms, also expressed in the lemmas of the base verbs. This factor is associated with the pervasiveness of participles, which indicates whether the participial word forms of a given verbal lemma are used more or less frequently than its corresponding finite and infinitival word forms. In the context of the exploratory analysis, pervasiveness reflects the preference of base verbs for deriving participial word forms instead of other finite and infinitival word forms in the corpus. The choice of these factors is not arbitrary. The adverb *očenʹ* is regarded as a formal marker of adjectivization that signals that a participle is adjectivized. Semantic classes (factor 2) reflect both the compatibility of verbal forms with *očenʹ* and the semantic affinity of a participle with its corresponding base verb; a participle derived from a base verb of a particular semantic class may convey a gradable semantic component that allows its use with the adverb *očenʹ*. Morphosyntactic factors (3) are based on the annotation of the manually disambiguated RNC subcorpus. There were no ambiguous readings (both adjective and participle) in this corpus, as the annotators had yet to disambiguate them as either participles or adjectives. The ambiguity tag conveying double reading ‘participle/adjective’ (or PTCP/ADJ) is provided by the analyzer, while the other tags are assigned by the annotators. For example, the word form *vydajuščijsja* ‘outstanding’ is tagged as *PrsAct* (present active participle) ‘sticking out’ and as *A* (adjective) ‘outstanding’ by the analyzer.

Exploring the distribution of the grammatical features (tense, voice, transitivity, and aspect) across the corpus frequencies of participles and their corresponding base verbs brings the following aspects to the fore. First, the analysis shows how the pervasiveness of participial forms in the corpus among other verbal (finite and infinitival) forms is related to the ambiguity and rank of their base verbs. Second, it establishes the relationship of morphosyntactic features of participles with the pervasiveness of participles and their ambiguity. The analysis is exploratory because I am primarily interested in patterns that are observed in the distributions of these factors, and address the following questions:

- How are these factors represented across corpus data?
- Can the claims of the syntactic and morphosemantic approaches be confirmed by corpus evidence? Are there any deviations from what is stated in the approaches and what is observed across the corpus data?
- How significant is the relationship of these factors with ambiguity?

The chapter is structured as follows. Section 2 focuses on analyzing the compatibility of participles and their corresponding finite verbal forms with the adverb of measure *očenʹ* ‘very’ (syntactic factor). It compares the constructions *očenʹ* + finite verbal form and *očenʹ* + participles and the semantic classes of the base verbs across their corpus frequency. The goal was to determine whether *očenʹ* only combined with adjectivized participles or adjectives only or if it could also combine with finite verbal forms and unambiguous participles. Section 2 also assesses the importance of compatibility with an adverb of measure/degree as a factor of adjectivization, and the role of the base verb semantics in this compatibility. Section 3.1 analyzes distributions of tense, voice, and transitivity features across the ratio of participles to their base verbs, ordered according to the ranks of their base verbal lemmas. The statistical analysis in Section 3.3 is a follow up to Section 3.1; the analysis predicts ambiguity expressed by the double reading that the analyzer assigns to a word form based on its morphosyntactic properties (tense, aspect, voice, and transitivity), annotated by the human experts in the RNC.

2 The adverbial *očen′* ‘very’ construction

In this section, I present an experiment concerning a specific case of the adjectivization factor, which is the compatibility of participles with the adverb of measure and degree *očen′* ‘very’ (also referred to as an intensifier). As a rule, *očen′* intensifies the qualitative meaning of adjectives. The intensifier can also be used with finite verbs with specific semantics. This renders the status of the intensifier as a factor of adjectivization somewhat conditional rather than definite because its use is extended to other verbal forms. For this reason, I needed to investigate the semantic properties of base verbs and syntactic contexts in which intensified verbal and participial word forms are used.

The adverb *očen′* characterizes the intensity of a given property, which is prototypically exhibited by an adjective or an adverb (Sičinava, 2018). *Očen′* can combine with verbs that have a gradable semantic component in their semantics. Sičinava (2018) states that the adverb can combine with *spešit′* ‘hurry’[1] as in *očen′ spešil* ‘[he] was in a great hurry’ (=‘acted quickly’), and with the verb *ljubit′* ‘love, like’[2] as in *očen′ ljubil* (=‘had a strong feeling’), but that it could not combine with the verb *idti* ‘go, walk’[3] as in **očen′ čël* ‘walked very’. Both *spešit′* ‘hurry’ and *idti* ‘go, walk’ have the semantic class of movement, but the former can be graded and therefore intensified, while the latter cannot. Thus, the use of *očen′* with finite verbal and participial word forms depends on whether the semantic class of this adverb is compatible with the semantics of verbs.

In this study, I address several objectives with regard to the *očen′* construction. First, I examine whether the gradable semantic component of the base verbs allows the use with *očen′* for participles and for finite verbal forms. I then identify which type of verb semantics and syntactic contexts can explain the intensification of participles. I also determine whether intensified participles lean towards adjectivization or if they remain unambiguous.

To answer these questions, I conducted an experiment on two types of corpus-based constructions, namely:

- *očen′* + finite verb
- *očen′* + participle

[1] A verb of movement.

[2] A verb of psychological domain.

[3] A verb of movement.

Each construction was represented by one verbal lemma. All the present finite verbal and participial word forms were grouped according to the lemmas of their corresponding base verbs. For example, the construction *očen′* + *podxodit′* corresponded to *očen′* + *podxodit* ‘ [. . .] fits-V.IND well’ and *očen′ podxodjaščij* ‘very befitting-PRESP’. The word form *podxodjaščij* can also be an adjectivized participle, ‘suitable’. Thus, each lemma was specified by a participial word form or a finite verbal word form in the indicative mood and the present tense only (without further specification of voice). Each *očen′* construction had the summed frequency of word forms grouped according to their corresponding lemmas.

First, I reviewed the corpus-based frequencies of lemmas for present indicative verbal and present participial lemmas. I then outlined the distribution of the ratios of the *očen′* constructions for participial and verbal word forms. Second, I studied the distribution of the semantic classes of the base verbs of the lemmas in these constructions. I also performed a statistical analysis of the distributions of semantic classes across the ratios of the constructions *očen′* + finite verbs and *očen′* + participles. Finally, I conducted a qualitative analysis of the syntactic contexts of the eight most frequent verbal and participial constructions. The aim of the analysis was to assess the strength of the association between the constructions and the ratio scores.

Corpus studies allow us to observe and analyze the participial and verbal constructions across their frequency ratios, the semantic classes of the base verbs, and the syntactic context in which they are used.

2.1 Data description

The corpus data used in this experiment were taken from Araneum Russicum Maius (Russia-only Russian, 15.03) 1.20 GB corpus,[4] which is part of the Aranea Project.[5] The corpus consists of texts in Russian in UTF-8 encoding, tagged using the MULTEXT-East Russian tag set,[6] and was compiled on 06/11/2019.

	Counts
tokens	1,200,000,258
words	859,319,823
sentences	71,616,173
paragraphs	29,510,308
documents	1,826,514

Table 4.2: Overview of counts in the Araneum corpus.

	Counts
word	8,114,574
lemma	7,075,180
tag	2,103

Table 4.3: The lexicon size of the Araneum corpus.

[4] Available at: http://unesco.uniba.sk/aranea_about/index.html

[5] The Aranea Project comprises a family of comparable Gigaword Web Corpora, prepared by Vladimir Benko within the framework of a joint project of UNESCO Chair in Plurilingual and Multicultural Communication (Comenius University in Bratislava) and Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences (Benko, 2014).

[6] Available at: http://nl.ijs.si/ME/V4/msd/html/msd-ru.html

The 1.20 GB version contains 1,200,000,258 tokens and 859,319,823 words, as shown in Table 4.2. Table 4.3 provides further information about the size of the lexicon in the corpus.

The *očen′* construction is based on present tense finite verbal and participial word forms. I chose these types of participles and verbs because they are less likely to be combined with *očen′* compared to the past tense word forms. In the Araneum corpus, the most frequent construction with *očen′* includes past tense participial word forms. Most of the word forms seem to be adjectivized and biased towards the use with *očen′*. For this reason, I used present tense participles (and their corresponding finite verbs), as they do not combine with *očen′* as frequently as do past tense participles.

2.2 Data analysis and interpretation

2.2.1 Overview of the distributions

I used queries based on the Corpus Query Language (CQL)[7] to retrieve the constructions and to separate word forms taken from the corpus:

- query 1: *[lemma="очень"] [tag="Vmip.*"]*
- query 2: *[lemma="очень"] [tag="Vmpp.*"]*
- query 3: *[tag="Vmip.*"]*
- query 4: *[tag="Vmpp.*"]*

The tag *Vmip* stands for main verb (*Vm*), indicative mood (*i*), and present tense (*p*). The tag *Vmpp* stands for main verb (*Vm*), participial word form (*p*), and present tense (*p*). All the present finite verbal and participial word forms were grouped under the present tense lemmas of their corresponding base verbs. This was done to visualize the correspondence between these constructions more easily and to refer to these lemmas when analyzing the semantic classes of the base verbs. Figure 4.1 illustrates the distribution of the present tense lemmas in the corpus. The present tense indicative verbs (referred to henceforth as PRS V) were about seven times more frequent than were the present participles (referred to henceforth as PRESP). The number of constructions with *očen′* appeared to be extremely low compared to the entire usage of verbal and participial lemmas. The constructions with present indicative verbs (*očen′* + PRS V) were 30 times more frequent than were the constructions with present participles (*očen′* + PRESP; 39.3 versus 1.2).

After retrieving the constructions, I compiled a list of the 90 most frequent verbal lemmas. Using this list, I set the queries as *očen′* + *list of lemmas* + [*tag="Vmip.*"*] or [*tag="Vmpp.*"*]. The complete list with the frequency and ratios of present participles, present indicative verbs, and the constructions *očen′* + PRESP and *očen′* + PRS V, is presented in Appendix D, Table D.1.

[7] CQL is a query language that is used to search for complex grammatical or lexical patterns, or other criteria that cannot be set using the standard user interface, available at: https://www.sketchengine.eu/documentation/corpus-querying/.

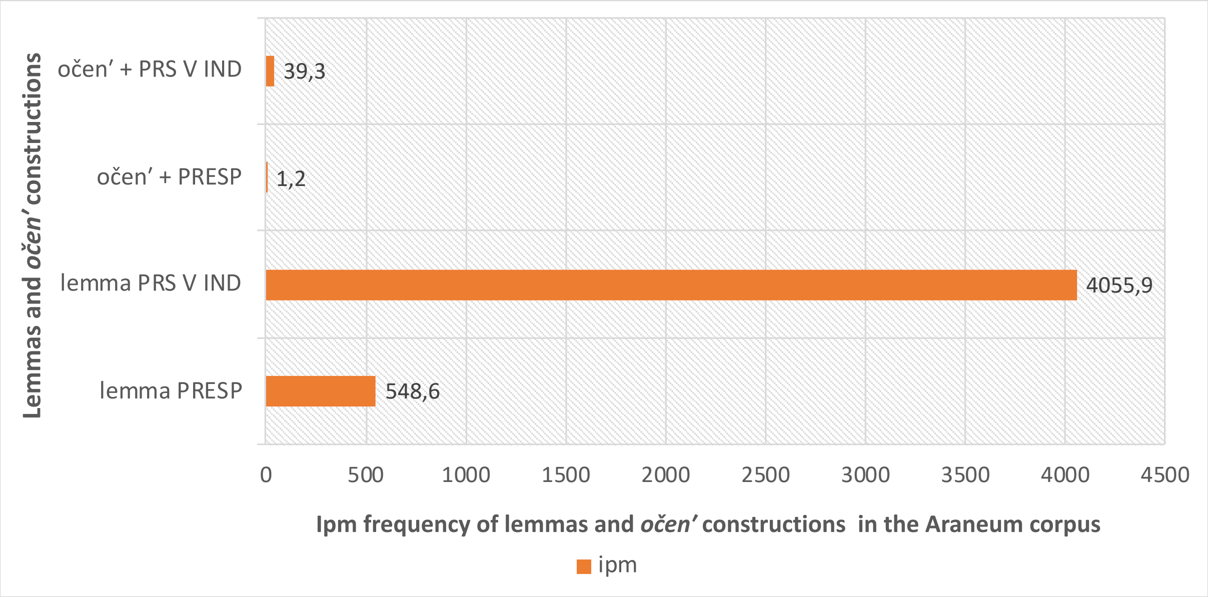


Figure 4.1: Distribution of lemmas and *očen′* constructions in the Araneum corpus.

Table 4.4 illustrates a sample of 22 *očen′* + PRESP and *očen′* + PRS V constructions with the ratio scores. The ratio for an *očen′* + PRESP construction was computed by dividing the number of its occurrences as a construction by the number of its occurrences as a present participle. The ratio for the *očen′* + PRS V construction was computed by dividing the number of its occurrences as a construction by the number of its occurrences as present indicative verbal word forms. The calculation of the ratio for the *očen′* + *stoit′* construction is provided below:

- ratio of *očen′* + PRESP construction = raw frequency of *očen′* + PRESP / raw frequency of PRESP = 2/8510 = 0.00024
- ratio of *očen′* + PRS V construction = raw frequency of *očen′* + PRS V / raw frequency of PRS V = 27/470393 = 0.00006

For the lemma *stoit′* ‘cost’, two present participles are only used with *očen′* versus 8,510 present participles with or without *očen′*, and 27 present tense finite verbal word forms are used with *očen′* versus 470,393 present tense finite verbal word forms (with or without *očen′*).

The ratios of the constructions were ordered according to the differences between the ratios of the participial and verbal constructions (from the lowest to the highest value), as shown in Table 4.7. The ratio of the *očen′* + PRESP (3rd column) and *očen′* + PRS V (4th column) constructions is given for each lemma (2nd column). In the interval of 46–56, the ratio of *očen′* + PRS V construction is higher than is the ratio for the *očen′* + PRESP construction. For example, for the lemma *goret′* ‘burn’ (#50), present indicative verbs (ratio of 0.00097) combine more often with *očen′* than do the corresponding present participles (ratio of 0.00037). In the interval of 57–67, the ratio of *očen′* + PRESP construction is higher than is the ratio for the *očen′* + PRS V construction.

# Rank	Verbal lemma	Ratio očen′ + PRESP	Ratio očen′ + PRS V
46	*umet′* ‘be able to’	0.00019	**0.00107**
47	*stremit′sja* ‘strive’	0.00017	**0.00097**
48	*idti* ‘suit, become’	0.00033	**0.00110**
49	*potrjasat′* ‘impresss’	0.00030	**0.00091**
50	*goret′* ‘burn’	0.00037	**0.00097**
51	*sootvetstvovat′* ‘correspond’	0.00015	**0.00067**
52	*vladet′* ‘possess, own’	0.00000	**0.00047**
53	*uznavat′* ‘recognize’	0.00000	**0.00038**
54	*obitat′* ‘live, reside’	0.00000	**0.00010**
55	*rabotat′* ‘work’	0.00001	**0.00008**
56	*pol′zovat′sja* ‘use’	0.00020	**0.00024**
57	*upravljat′* ‘govern, manage’	**0.00009**	0.00002
58	*trebovat′* ‘require’	**0.00007**	0.00003
59	*oxranjat′* ‘guard’	**0.00007**	0.00000
60	*proxodit′* ‘pass’	**0.00013**	0.00005
61	*stoit′* ‘cost’	**0.00024**	0.00006
62	*vyzyvat′* ‘call’	**0.00022**	0.00002
63	*blestet′* ‘shine’	**0.00394**	0.00368
64	*vozbuždat′* ‘excite, agitate’	**0.01325**	0.01292
65	*ponimat′* ‘understand’	**0.00477**	0.00433
66	*dumat′* ‘think’	**0.00055**	0.00002
67	*ožidat′* ‘await’	**0.00111**	0.00047

Table 4.4: A sample of 22 *očen′* + PRESP and *očen′* + PRS V constructions and their ratio values. The largest ratio value in each row is in bold.

2.2.2 Overview of the semantic properties of the lemmas in the most frequent constructions

In this section, I discuss the semantic properties of the lemmas used in the *očen′* construction. I annotated 90 lemmas using Kustova’s (2001) database of 19,583 verbal lemmas and their semantic classes; for example, *t:perc* (perception), *t:move:body* (movement), *t:be:exist* (existential), and so on.[8] Some verbs in the list of 90 lemmas were missing in Kustova’s database, which is why I annotated them myself (see the full list of the semantically annotated lemmas in Appendix D, Table D.2) or by referring to the Russian FrameBank;[9] see Lyashevskaya and Kashkin (2015, 2014). Table 4.5 illustrates the list of the semantic classes used for *očen′* + PRESP and *očen′* + PRS V.

I divided the list of the verbal lemmas into two intervals, with 1–56 verbal lemmas covering the cases in which *očen′* + PRESP were less frequent than *očen′* + PRS V, and 57–90 lemmas for which *očen′* + PRESP were more frequent than *očen′* + PRS V. I then computed the percentages of the semantic classes per interval.

[8] Galina Kustova’s database of semantic classes of verbal lemmas is part of the semantic classification in the RNC. For the complete description of the semantic classes in the RNC, see http://www.ruscorpora.ru/en/corpora-sem.html

[9] Available at: https://github.com/olesar/framebank

Semantic class	Description
be:exist	existence
changest	change of state or property
contact	physical contact and support
impact	physical effect
impact:creat/be:creat	creation of a physical object/creation in the domain of existence
light	light (e.g., beam)
ment	mental domain
move	movement
perc	perception
physiol	physiological domain
poss	domain of possession
psych	psychological domain
sound	sound (e.g., noise)
speech	speech (e.g., discussion)

Table 4.5: Semantic classes used for the more frequent verbal constructions and participial constructions.

Figure 4.2 illustrates the percentages of semantic classes for the interval 1–56.

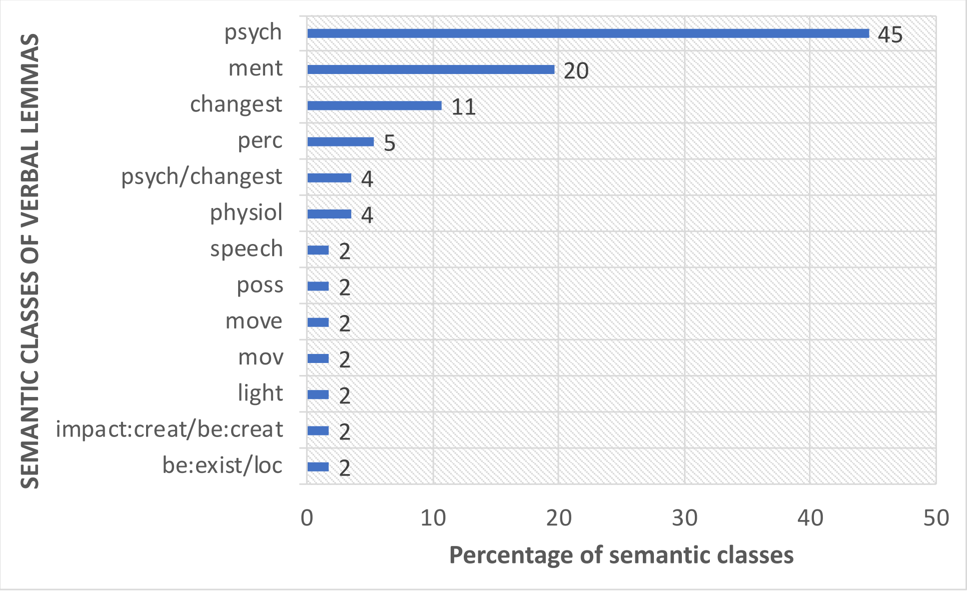


Figure 4.2: The percentages of the semantic classes for the interval with more frequent *očen′* + PRS V constructions (1–56).

The most frequent classes denote the psychological domain (*psych* accounts for 45%; for example, *nravit′sja* ‘like’), the mental domain (*ment* accounts for 20%; for example, *somnevat′sja* ‘doubt’), and the change of state or features (*changest* accounts for 11%; for example, *razvivat′* ‘develop’). Overall, many verbs are related to mental, psychological, and perception domains, which are likely to be gradable and intensified by the adverb *očen′*.

Figure 4.3 shows the percentages of the semantic classes for the interval 57–90. The most common classes are the mental domain at 24% (for example, *ponimat′* ‘understand’), the psychological domain at 21% (for example, *uvlekat′sja* ‘be fond of’), and the domain of speech at 15% (for example, *govorit′* ‘talk’). Other verbs have semantics that are related to the sensory domain (*sound*, *contact*, and *physiol*),[10] but the number is not as representative in comparison to the classes with percentages of over 10%.

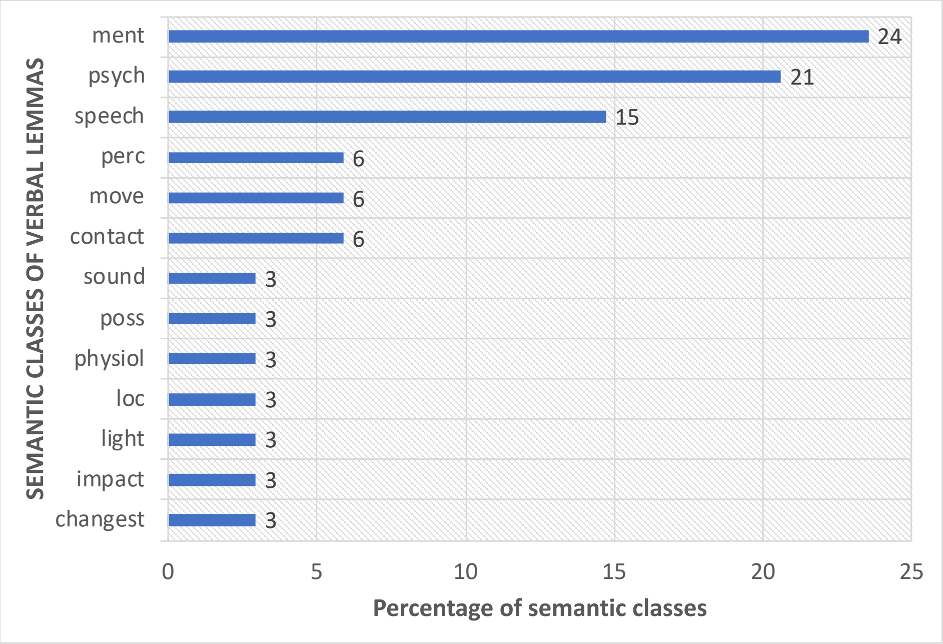


Figure 4.3: The percentages of the semantic classes for the interval with more frequent *očen′* + PRESP constructions (57–90).

2.2.3 Statistical analysis of the ratio distributions and semantic classes

In this section, I assess the statistical significance of the observations concerning the ratio distributions of the constructions and the semantic classes of the base verbs. I investigate whether there is a significant difference in the ratios of *očen′* + PRESP and *očen′* + PRS V constructions. I also estimate the strength of the association between the ratios of the constructions and the semantic classes. The assessment of the statistical differences in the ratios of the distributions of the verbal and participial constructions, as well as of the strength of association between the ratios and the semantic classes, allow for the observation of the following: (a) whether participles or finite verbs are more likely to be used with *očen′*, (b) whether the semantic classes of the base verbs affect the ratios of the verbal and participial constructions, and (c) which semantic classes have the strongest effect on the ratios in *očen′* constructions. The complete list of commands used to compute the scores and coefficients for the constructions is presented in Appendix D, Section E.

[10] For example, *xrustet′* ‘crunch’ for *sound*, *oblegat′* ‘fit tightly’ for *contact*, and *pit′* ‘drink’ for *physiol*.

In order to test the significance of the observations, I devised the following hypotheses:

- *Null Hypothesis* H_{01}: The ratio distributions of *očen′* + PRESP and *očen′* + PRS V constructions are the same.
- *Null Hypothesis* H_{02}: The ratio of *očen′* constructions among the semantic classes is the same.
- *Hypothesis* H_1: The ratio distributions of *očen′* + PRESP and *očen′* + PRS V constructions are significantly different.
- *Hypothesis* H_2: The ratio of *očen′* constructions among the semantic classes is significantly different.

I used the same dataset containing the ratio distributions for 90 lemmas (N = 90). The dataset consisted of two continuous variables, *ratioptcp* and *ratiov* (standing for the ratios of the participial and verbal constructions), and one categorical variable, *semrole* (standing for the semantic classes of the base verbs). The initial number of levels[11] (n = 16) in *semrole* was reduced to six by comparing the averaged values of each level and replacing less frequent levels with a new level *other*. Among the reduced levels, *psych*, *other* and *ment* were the most frequent; *other* was a new level grouping the less frequent levels. Table 4.6 illustrates the levels and their counts.

Semantic classes	**n**
psych	35
other	21
ment	16
changest	6
perc	6
speech	6

Table 4.6: Counts of semantic classes.

Although I removed the outliers from *ratioptcp* and *ratiov* prior to visualization and statistical tests, the variables *ratioptcp* and *ratiov* were both non-normally distributed, with a skewness of 1.90 and 0.75, respectively. This indicates that the ratio distributions were skewed towards the left, and that non-parametric tests were needed for testing the hypotheses. The average for *ratioptcp* was 0.005 (standard deviation SD = 0.008); for *ratiov*, it was 0.010 (SD = 0.011).

Box plots that illustrate the differences between the semantic classes for the ratio distribution of the construction *očen′* + PRESP are shown in Figure 4.4. The box plots (*changest*, *psych*, and *other*) appear to be right-skewed (as there is a wider range in the values in the upper quartile of these box plots), while *perc* is left-skewed because there is a wider range in the values in the lower quartiles. The box plots *ment* and *speech* have an approximately symmetric distribution – that is, the lower and the upper quartiles in these box plots have almost the same number of data

[11] Levels are categories of the semantic classes in the *semrole* variable, such as *psych*, *ment*, *changest*, *perc*, and so on.

points. These observations imply that the *očen′* constructions tended to have a higher ratio when they were used with the participles that derived from the base verbs of the mental domain, the psychological domain, and *other* (for example, *possession*, *sound*, and *movement*).

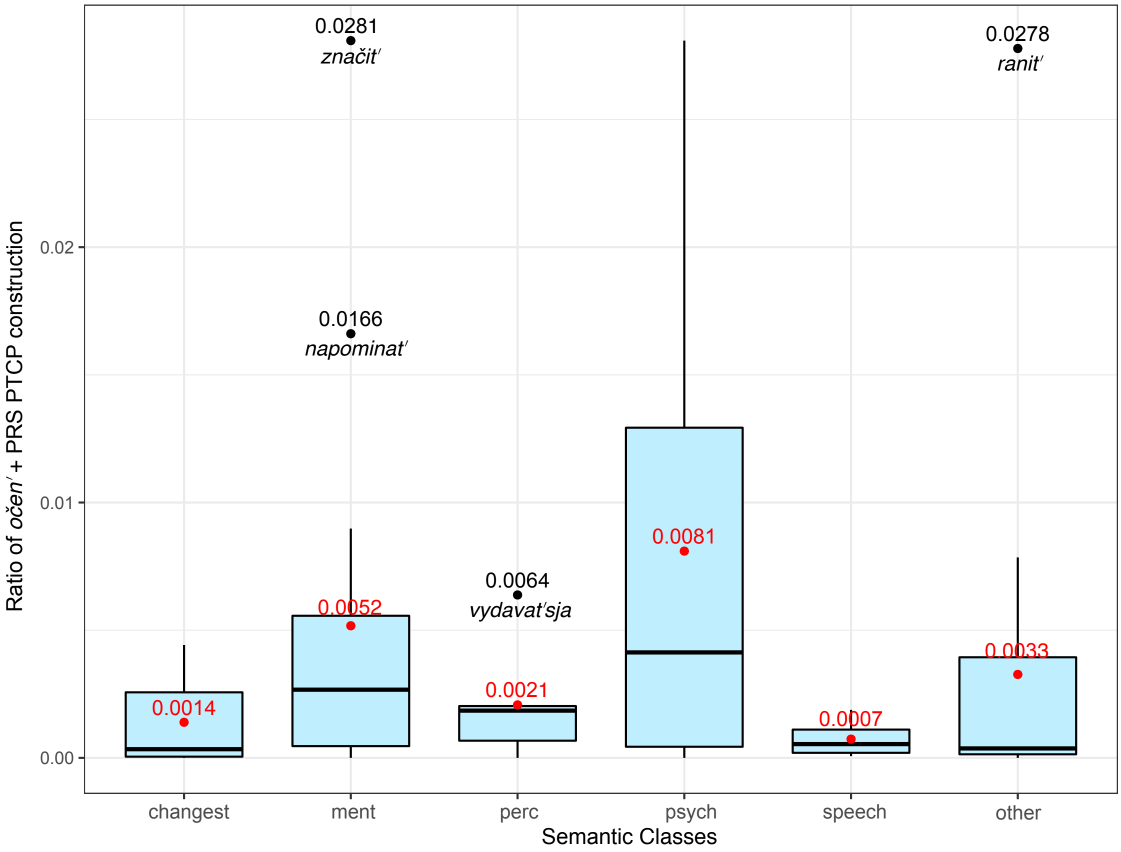


Figure 4.4: The semantic classes of the base verbs across the ratio distributions of the constructions *očen′* + PRESP.

The outliers for the semantic class *ment* are *značit′* ‘mean’ and *napominat′* ‘remind, resemble’ (ratios of 0.0281 and 0.0166, respectively), for *perc*, *vydavat′sja* ‘stick out, be prominent’ (a ratio of 0.0064) and for the class *other*, *ranit′* ‘hurt’ (a ratio of 0.0278). The examples for *značit′* ‘mean’ include short present passive forms, such as *ochen znachim* ‘very important-ADJ’ and one occurrence of a full present passive form, *ochen znachimye* ‘very important-ADJ’. Most of these word forms seem to be adjectivized and have undergone a shift in meaning. For this reason, the lemma *značit′* for from the class *ment* has a higher ratio in the *očen′* construction compared to the other lemmas from the class *other*. The only two attested examples of the outlier *ranit′* ‘hurt’ are the present passive forms *ranimye*, which now convey the qualitative meaning of ‘vulnerable’, extended from the verbal meaning ‘[being] hurt’.

Mean (*M*) scores and their positions in the box plot are shown in red. The box plot *psych* has the highest average frequency (*M* = 0.0081), followed by *ment* (*M* = 0.0052) and *other* (*M* = 0.0033). The mean for the rest of the box plots is below 0.003. The frequency values of the participial constructions vary most for *psych*, followed by *ment*, as these box plots are the tallest ones. The box plots *speech* and *perc* are the shortest ones compared to the others. The short box plots suggest that the distributions of the ratios for the *speech* and *perc* classes are the least dispersed and lack extreme values. The low dispersion for these classes is explained by the small number of their corresponding lemmas (six), while the high dispersion for *psych*, *other*, and *ment* is explained by the larger number of lemmas (35, 21, and 16), as shown in Table 4.6. There are only two outliers in *ment*, one in *perc*, and one in *other*. The tails indicating the extremities of the frequency values are the longest in *psych*, *other*, and *ment*. Thus, participles derived from the verbs in the psychological and mental domains tended to be used more frequently in the *očen′* construction compared to the participles derived from the verbs in the other semantic classes.

Box plots that show the differences between semantic classes for the ratio distributions of the constructions *očen′* + PRS V are presented in Figure 4.5. The box plots *changest* and *psych* appear to be almost normal with a slight right skew, while the rest of the box plots are right-skewed. The observations with *ment*, *speech*, and *other* contained one to two outliers. The box plot *speech* is underrepresented because its lower quartile is equal to the minimum as a result of the low number of counts (see Table 4.6). The data are mainly dispersed in *changest* and *psych*, and show the highest mean ratio for lemmas (for example, M = 0.0132 for *changest*, and M = 0.0168 for *psych*) in these box plots. The extremities of the frequency values are the greatest in *psych*, *changest*, and *ment*.

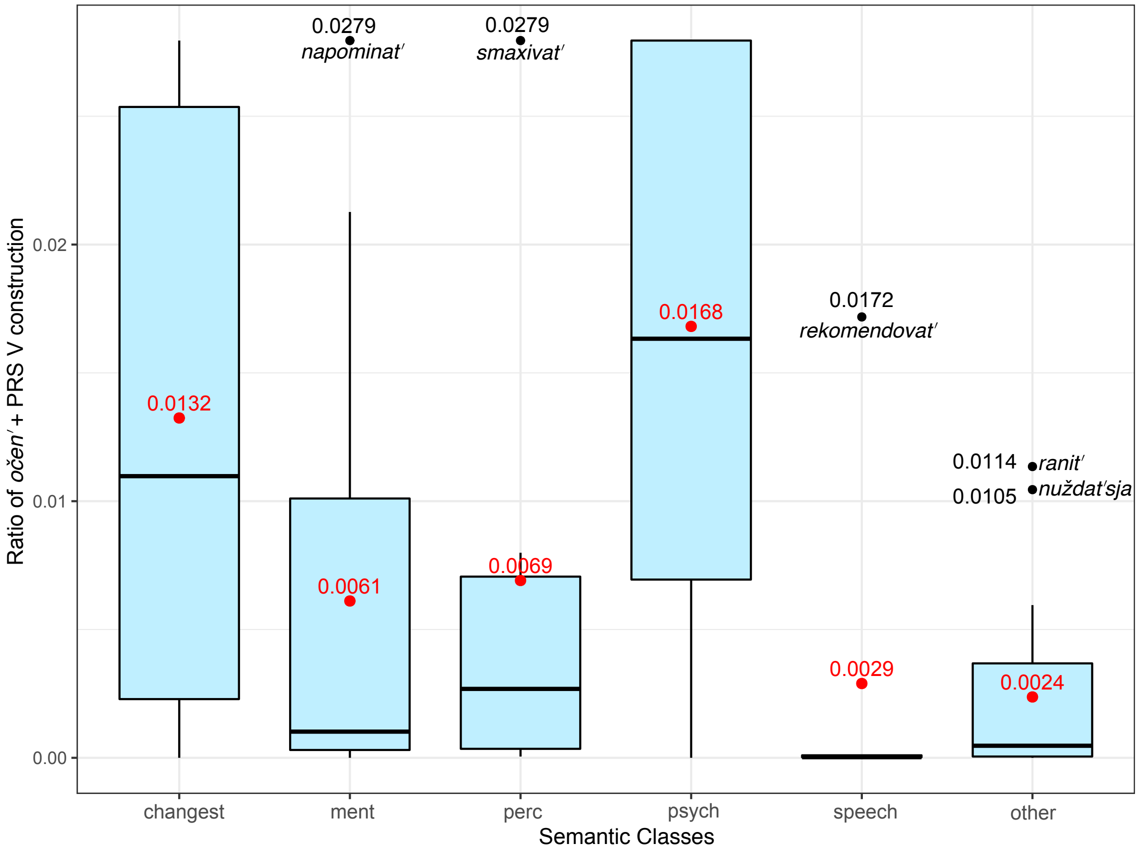


Figure 4.5: The semantic classes of the base verbs across the ratio distributions of the constructions *očen′* + PRS V.

Overall, the graph shows that the verbal forms with the semantic classes of change of state and psychological domain were used more frequently in the *očen′* construction compared to the classes of mental domain, perception, speech, and others. There were five outliers for the four semantic classes, namely the verb *napominat′* 'remind, resemble' for *ment*, *smaxivat′* 'look like' for *perc*, *rekomendovat′* 'recommend' for *speech*, and *ranit′* 'hurt', *nučdat′sja* 'need' for the remaining semantic class *other*. The lemma *ranit′* 'hurt' was an outlier that was found in both the participial and verbal constructions due to its exceedingly high ratio (0.0278 and 0.0114, respectively) compared to the ratios of the remaining lemmas in *other*.

The box plots in Figure 4.4 illustrate that participles were used less frequently with *očen′* than finite verbs were used (*cf.* Figure 4.5). Moreover, the semantic classes of psychological, mental domains, and change of state appeared to be prominent both in the distribution of verbal and participial constructions. The data displayed in the box plots are independent[12] and not normally distributed; it also consists of continuous (*ratioptcp* and *ratiov*) and categorical (*semrole*) variables. These criteria enabled the use of non-parametric techniques for assessing the strength of the association between the distributions of verbal and participial constructions with their corresponding semantic classes. The tests, which suited the hypotheses stated at the beginning of Section 2.2.3 and matched non-normal distributions of the *očen′* constructions, were non-parametric Wilcoxon rank-sum, Kruskal–Wallis, and Dunn's tests[13](*cf.* Wilcoxon, 1945; Kruskal and Wallis, 1952; Dunn, 1961). Lijffijt et al. (2014) and LaVange and Koch (2006) discuss the use of non-parametric tests as an appropriate (especially for analyzing frequency distributions of words) and robust alternative to parametric techniques. Lijffijt et al. (2014) argue that the Wilcoxon rank-sum test, among other non-parametric tests, makes the fewest assumptions about the frequency distribution and is thus the most generally applicable for analyzing frequency distributions of words drawn from multiple texts. Testing of words with sufficient frequency and/or dispersion using the Wilcoxon test also yields proper results (Lijffijt et al., 2014: 385). When the data do not meet one or more of the assumptions for parametric procedures (e.g., DePoy and Gitlin, 2016), non-parametric methods will demonstrate good power properties, with statistical power equal to 93% of that expected for standard parametric methods (such as *t*-tests) applied under ideal circumstances (LaVange and Koch, 2006: 2532).

First, I used to the non-parametric Wilcoxon rank-sum test (also referred to as the independent 2-group Mann-Whitney U Test) with a continuity correction. The Wilcoxon rank-sum test is used to prove a significant difference between two sample groups using magnitude-based ranks. It compares two distributions obtained between two separate groups to assess whether one has systematically larger (and therefore different) values compared to the other[14] (Haynes, 2013;

[12] This property refers to independence of observations, which means that there is no relationship between the observations in each distribution or between the distributions themselves.

[13] The tests are computed with the R functions: *wilcox.test*(*y,x*), *kruskal.test*(*y~x*) and *dunn_test*(*y~x*) (R Core Team, 2019). R is an open source programming language and environment for statistical computing, available at: http://www.R-project.org

[14] If the ranks of the two sample groups are significantly different, the test statistic identifies a significant difference.

Field, 2013). In this test, I compared the continuous variables *ratioptcp* and *ratiov*; that is, the ratio distributions of the participial and verbal constructions with *očen′*. The test revealed that the p-value[15] was lower than 0.05 (p = 0.035), which means that there was a significant difference between the ratios of the participial and verbal constructions with regard to whether they were used more or less frequently with *očen′*. The p-value of the Wilcoxon test allowed me to reject H_{01} and to confirm H_1.

Since the ratio distributions were non-normal, I conducted a non-parametric Kruskal-Wallis test to examine the differences in the ratios of the two types of constructions with regard to the semantic classes of the base verbs. The Kruskal-Wallis test is a rank test that maintains the assumption of independent random samples from each population, and can be used when the assumption of normality is violated. This test is based on the sum of the ranks for the groups that are compared.[16] The test was run on (a) the dependent continuous variable *ratioptcp* and the independent categorical variable *semrole*, and (b) the dependent continuous variable *ratiov* and the independent categorical variable *semrole* (the same as in (a)). Significant differences were found among all the six semantic classes in the ratios of the participial constructions (chi-square χ^2 = 11.37, degree of freedom *df* = 5, p = 0.045), and in the ratios of the verbal constructions (χ^2 = 32.732, *df* = 5, p < 0.001). The p-value for the verbal constructions was considerably more significant than was that for the participial constructions. This implies that the semantic classes of the base verbs were associated less strongly with the ratio of participial constructions compared to the ratio of the verbal constructions.

Since the Kruskal-Wallis test was significant, I conducted a post-hoc[17] Dunn’s test with a Bonferroni adjustment[18] of p-values, which is used to compare distinct groups at the baseline and is conducted on each pair of groups. More specifically, the test estimates pairwise comparisons between the groups of independent variables; that is, semantic classes such as *changest–ment*, *changest–perc*, and so on.[19] Again, the dependent variables were *ratioptcp* and *ratiov*; and the independent variable was *semrole*. The pairwise comparisons for *ratioptcp* differences among all the semantic classes were not significant, while the comparisons for *ratiov* were significant for the classes *ment* and *psych* (p = 0.014), *psych* and *speech* (p = 0.004), and *psych* and *other* (p < 0.001). The z-scores of the Dunn’s test for *ratiov* also showed that *psych* was more stochastically dominant than was *ment* (z-score of 3.31) in the group *ment–psych*, while *psych* was statistically different from *speech* and *other* in terms of higher ratios (z-scores of -3.66 and -4.87, respectively) in the groups *psych–speech* and *psych–other*. No significant differences were found in the rest of

The calculations use the simple addition of ranks, and significance is based on an established distribution.

[15] The p-value of the Wilcoxon test is based on the sampling distribution of the rank sum statistics, in which the null hypothesis (no difference in distributions) is true.

[16] The more different the sums, the stronger the evidence that the responses (that is, the dependent variables) are systematically larger in some groups than in others (Field, 2013).

[17] A post hoc test is also referred to as a ‘multiple comparison test’ (MCT). Post hoc tests are used to identify the specific differences between sample pairs in order to analyze specific sample pairs for significant difference(s).

[18] The Bonferroni adjustment multiplies each Dunn’s p-value by the total number of tests being conducted.

[19] See Appendix E for the complete output of the pairwise comparisons of the groups representing the semantic classes.

the classes in *ratiov*. Thus, the significant scores of the Kruskal-Wallis and Dunn's tests allowed me to reject H_{02} and accept H_2 for the ratio distribution of the verbal constructions. H_2 for the ratio distribution of the participial constructions, given the significant p-values of the Wilcoxon and Kruskal-Wallis tests, was accepted, although multiple comparisons among the semantic classes in the Dunn's test showed no significant differences. The latter might be explained by the fact that the Dunn test was somewhat conservative for the small values of the semantic classes in *ratioptcp*.[20] It may also imply that participles bound by semantic classes show a less clear pattern in term of compatibility with *očen′*.

The results of the statistical analysis showed the following tendencies. The semantic class of psychological domain appeared to be prominent in the participial and verbal constructions. It contributed to higher ratios of *očen′* used with verbal constructions, while it applied to both higher and lower ratios in the participial constructions. The verbs of mental domain represented the second most relevant class (associated with a higher ratio of *očen′*) in the distribution of the participial constructions, as did the verbs of change of state in the distribution of the verbal constructions. The distributions in the participial and verbal constructions differed significantly in their ratios, also based on the semantic classes.

Based on the significantly lower mean values in the box-plots in Figure 4.4, it is somewhat unlikely for a participial word form to combine with *očen′* (also in comparison to a finite verbal word form). While the finite verbal forms of psychological and mental domains used with *očen′* differed significantly in their ratios in other verbal forms, none of the participial forms used with *očen′* and belonging to a particular semantic class showed a significant difference.

2.2.4 Context analysis of the lemmas in the most frequent constructions

In this section, I discuss the eight most frequent *očen′* + PRESP and *očen′* + PRS V constructions. The objective is to qualitatively examine the syntactic context of these constructions and to identify any uses of adjectivized participles. More specifically, I examine some syntactic factors of adjectivization in these constructions, and investigate whether *očen′* (as one of the factors) combines with the finite verbal forms the same way it does with participial, ambiguous, and unambiguous word forms.

I searched manually for examples that contained the finite verbal and participial word forms, and investigated whether they were ambiguous or unambiguous by analyzing their context and meaning. I then randomly selected one example of a construction with a finite verbal form, and one with a both ambiguous and unambiguous participial form. I selected the eight most frequent constructions (*očen′* + PRESP and *očen′* + PRS V) and placed their ratios side by side in increasing order of the *očen′* + PRESP ratio. Due to the high frequency of the *očen′* + PRESP constructions, there were many examples with both unambiguous and adjectivized forms of

[20] The small values concern the lower dispersion of lemmas and the lower ratio values in the box plots of the semantic classes for the participial constructions.

participles compared to the less frequent participial constructions used with *očen′*.

Table 4.7 illustrates the ratios of eight lemmas grouped by participles in *očen′* + PRESP and finite verbs in *očen′* + PRS V constructions. For the first four lemmas (*ljubit′* ‘love’, *napominat′* ‘remind’, *podxodit′* ‘suit’ and *idti* ‘suit, become’), the ratio of *očen′* + PRS V is higher than is that of *očen′* + PRESP. For the other four lemmas (*znat′* ‘know’, *zapominat′sja* ‘remember’, *zaxvatyvat′* ‘captivate, fascinate’, and *značit′* ‘mean’), the ratio of *očen′* + PRESP is higher than is that of *očen′* + PRS V. Most of the lemmas share the semantic classes of psychological, mental, and perception domains. The extended table, including the raw frequencies for constructions and single word forms, is provided in Appendix D, Section 3 (Table 4.7).

Verbal lemma	**Semantic class**	**Ratio *očen′* + PRESP**	**Ratio *očen′* + PRS V**
ljubit′ ‘love’	psych	0.0155	**0.0988**
napominat′ ‘remind, resemble’	ment	0.0166	**0.0347**
podxodit′ ‘suit’	ment	0.0040	**0.0099**
idti ’suit, fit, become’	perc	0.0003	**0.0011**
znat′ ‘know’	ment	**0.0022**	0.0001
zapominat′sja ‘remember’	ment	**0.0079**	0.0042
zaxvatyvat′ ‘captivate, fascinate’	psych	**0.0126**	0.0068
značit′ ‘mean’	ment	**0.0702**	0.0000

Table 4.7: A sample of the eight *očen′* + PRESP and *očen′* + PRS V constructions with the highest frequency among the other constructions in the entire ratio distribution. The largest ratio value in each row is in bold, and the decimals are rounded to four places.

2.2.4.1 Constructions with more frequent finite verbal forms

The example in (1a) illustrates that *očen′* combines with the finite verbal form *ljublju* ‘[I] love’ (a verb of psychological domain). In (1b), the present passive participial form *ljubima* ‘loved’ is used with the agentive instrumental complement *svoej xozjajkoj* ‘her master’, in (1d), the present passive participial form *ljubjaščaja* ‘loving’ joins the complement with the infinitival form *poest′* ‘to eat’. (1c) stands apart from the rest of the examples as (a) it does not use complements or adjuncts, and (b) the meaning of *ljubjaščim* ‘loving’ no longer characterizes the verbal action, but modifies the head noun *synom* ‘son’. In addition, the use of *očen′* with this word form implies that it is adjectivized because it intensifies the qualitative attributive meaning of *ljubjaščim*. Given that this word form is also found as an adjective in Zaliznjak’s (2003) dictionary, I suggest that *ljubjaščim* should be adjectivized, and that *očen′* should intensify its adjectival, and not its verbal meaning.

(1) lemma *ljubit′* ‘love’

a. *Ja* ***očen′ ljublju*** *tjažëluju muzyku (xard, metall i pročee).*
I very love.PRS.3SG heavy music hard metal and other

‘I am fond of heavy music (hard, metal and other).’

b. *sobačka živa i zdorova, i daže* ***OČEN′ ljubima*** *svoej xozjajkoj*
dog alive and healthy and even very love:PP her master.INS.SG
‘this dog is feeling well and even much loved by her master’

c. *Oleg vsegda byl* ***očen′ ljubjaščim*** *synom.*
Oleg always was very loving.INS.SG.M son
‘Oleg has always been a very loving son.’

d. *Èto šustraja trexcvetnaja košečka,* ***očen′ ljubjaščaja*** *poest′.*
this agile three.year cat very love:PRESP eat.INF
‘This is an agile 3-year-old pussy cat, [much] loving to eat.’

With regard to the lemma *napominat′* ‘remind, resemble’ (verb of mental domain), I only found unambiguous verbal forms.[21] The adverb *očen′* combines with the present indicative form *napominaet* ‘resembles’ in (2a), and with the present participle *napominajuščie* ‘resembling’ used with the verb complement *raz″em* ‘outlet’ in (2b).

(2) lemma *napominat′* ‘remind’

a. *Takaja dieta* ***očen′ napominaet*** *razgruzočnyj den′.*
Such diet very resemble.PRS.3SG fasting day
‘Such diet strongly resembles a fasting day.’

b. *mini USB,* ***očen′ napominajuščie*** *raz″em, ispol′zuemyj Olimpusom*
mini USB very resemble:PRESP outlet.ACC.SG used Olympus
‘[outlets on the camera] are mini USBs resembling a lot the outlet used in Olympus[22]’

In (3a), *očen′* intensifies the finite verbal form *podxodit* ‘fits’ (verb of perception). In (3b), the present participle *podxodjaščaja* ‘well suited’ is used with the adjunct *dlja poletov* ‘for flights’. Although the adjunct is a factor of adjectivization, the example in (3b) is borderline because the word form *podxodjaščaja* can be substituted by a synonymous adjective, *udačnyj*‘suitable’. (3c) shows that the word form *podxodjaščee* ‘suitable’ is adjectivized because it is not used with adjuncts or complements, is intensified by *očen′*, and conveys the meaning that qualifies *vremya* ‘time’ as suitable. In this example, *podxodjaščee* combines with *očen′* due to its adjectival meaning, which can be intensified.

(3) lemma *podxodit′* ‘suit, fit’

a. *Dlja menja kreslo* ***očen′ podxodit,*** *t.k. v nëm udobno rabotat′.*
for me armchair very suit.PRS.3SG as in it comfortable work
‘The armchair suits me a lot, as it is comfortable to work in it.’

b. *Pogodka to ne* ***očen′ podxodjaščaja*** *dlja poletov.*
weather not very suit:PRESP for flights
‘Looks like this weather is not well suited for flights.’

[21] This lemma was also an outlier for both present participles and finite verbal forms.

[22] In this sentence, Olympus refers to a camera brand.

c. *nesmotrja na, kazalos′ by,* ***očen′ podxodjaščee*** *vremja*
despite on seemingly very suit:presp/adj time
'despite the seemingly very suitable time'

The examples in (4a) and (4b) show that the lemma *idti* 'suit, become' (verb of perception) is only used with *očen′* in the figurative meaning of 'suit, fit, become (of clothing)' instead of 'go'. Among the occurrences with *očen′*, there were no examples of ambiguous participles for the lemma *idti*. In (4b), the participle *iduščaja* 'becoming' is used with the dative complement *mne* 'me', thus preserving its verbal properties, despite the fact that it is used with the figurative meaning of the main verb *idti*.

(4) *idti* 'suit, become'

a. *Ja sčitaju, čto ko mne.* ***očen′ idut*** *krasnyj, sinij i belyj*
I think that to me.dat.sg very become.prs.3pl red blue and white
cvet.
color
'I think that red, blue, and white become me well.'

b. *u menja byla . . . stil′naja i* ***očen′ iduščaja*** *mne pričeska*
at me was . . . stylish and very become:presp I.dat.sg hairstyle
'I had [the most refined,] stylish and very becoming [to me] hairstyle'

The observations show that there are both adjectivized and unambiguous participial word forms in the participial constructions with lower ratios (compared to the finite verbal constructions with *očen′*). Unambiguous participles are used with adjuncts and complements, while adjectivized ones stand alone. There is no difference in meaning in the finite verbs and (adjectivized/unambiguous) participles used with *očen′*. For example, the figurative abstract meanings of the base verbs *napominat′* 'remind, resemble' and *podxodit′* 'suit, fit', *idti* 'suit, become' remains the same as in their corresponding participles. I suggest that verbal word forms are used more often with *očen′* than with participial forms because participles retain the same semantics as finite verbs, and there is considerable occurrence of present indicative verbs in comparison to present participles in the Araneum corpus (Figure 4.1). In addition, the examples of adjectivized participles related semantically to their base verbs show that the lack of semantic relatedness does not always hold as a distinct factor of adjectivization.

2.2.4.2 Constructions with more frequent participial word forms

Most of the finite verbal constructions derived from *znat′* 'know' (verb of mental domain) are used with the negation particle *ne* 'not', as in *ne očen′ znaju* '[I] do not know well'; see (5a). One of the exceptions is provided in (5b), in which the word form *znajut* '[they] know' is used (without the negation particle) as part of the set expression *znat′ tolk* 'be a good judge of'. In (5c), *znajuščij* is used as an adjectivized participle;[23] that is, without adjuncts or complements, and in

[23] This also applies to the rest of the examples for participial forms.

a preposed position to the head noun *čelovek* 'person' that it modifies. In addition, it is used as part of a sequence with another adjective, *porjadočnyj* 'decent'. I did not find any occurrences of unambiguous participles used in this construction in the Araneum corpus.

(5) *znat′* 'know'

a. *ja . . . ne prepodaju, tak čto ne* ***očen′ znaju*** *nynešnjuju situaciju*
I . . . not teach so not very know.PRS.1SG current situation
'I haven't taught [for 10 years in Russia] so I do not know the current situation well.'

b. *naši ljudi tolk v nastojaščix veščax daže* ***očen′ znajut***
our people sense in genuine things even very know.PRS.3PL
'our people are even very good judges of genuine things'

c. *Èto i* ***očen′ znajuščij,*** *i očen′ porjadočnyj čelovek.*
this and very knowledgeable.NOM.SG.M and very decent person
'This is both a very knowledgeable and decent person.'

In (6a), the finite reflexive verbal form *zapominaetsja* 'is memorized' that is used with *očen′* has the semantic class of psychological verbs. In (6b), *zapominajuščimsja* 'memorable' is likely to be adjectivized because (a) it is intensified by *očen′*, (b) it is used in a row with another adjective *bogatym* 'rich', as in *bogatym na jarkie vpečatlenija* 'rich in vivid impressions', and (c) it can be substituted by the synonymous adjective *jarkim* 'bright'. There were no other examples of clearly unambiguous participles intensified by *očen′*.

(6) lemma *zapominat′sja* 'remember'

a. ***očen′ zapominaetsja*** *muzyka, s kotoroj načinaetsja . . . serial*
very memorize.PRS.REFL.3SG music from which begins . . . series
'the music at the beginning [and the end] of the series is easy to memorize'

b. *2014 god byl* ***očen′ zapominajuščimsja*** *i bogatym na jarkie vpečatlenija*
2014 year was very remember:PRESP/ADJ and rich for vivid impressions
'the year 2014 was very memorable and rich in vivid impressions'

I found no examples with unambiguous participles of the lemma *zaxvatyvat′* (verb of the psychological domain). (7a) illustrates the use of the finite verbal form *zaxvatyvaet* 'fascinates' with *očen′*, while (7b) shows the use of the ambiguous present participle *zaxvatyvajuščee* 'exciting'. This word form appears to be adjectivized, as it is used without adjuncts and complements, and modifies *zanjatie* 'activity' as being exciting. Therefore, in (7b), *očen′* intensifies the qualitative adjectival, rather than the verbal, meaning.

(7) lemma *zaxvatyvat′* 'captivate, excite'

a. *gejmplej . . .* ***očen′ zaxvatyvaet,*** *osobenno detej.*
gameplay . . . very fascinate.PRS.3SG especially children
'the gameplay [is made similar to Mario games and] fascinates you a lot, especially children.'

b. *Poisk klada dejstvitel′no,* ***očen′ zaxvatyvajuščee*** *zanjatie, i azartnoe!*
search treasure indeed very excite:PRESP/ADJ activity and adventurous
‘Searching for treasure is indeed a very exciting activity, as well as adventurous!’

I also did not find any examples of unambiguous participles for the lemma *značit′* ‘mean’ (verb of mental domain), only the finite verbal (as *značit* ‘matters’ in (8a)) and ambiguous participial forms (as in (8b)). The example in (8b) shows that *značim*, as a short form of *značimyj* ‘significant’, combines with *očen′*, is followed by the adjunct *dlja blagopolučija* ‘for the well-being’, and is used predicatively. The predicative use and the presence of an adjunct indicate that the word form is participial; however, the meaning of *značim* has shifted from the meaning of the main verb *značit′* ‘mean’, and now conveys the meaning of ‘significant, important’.

(8) lemma *značit′* ‘mean’

a. *Vaš opyt* ***očen′ značit****.*
your experience very mean.PRS.3SG
‘Your experience matters a lot.’

b. *Trud ètix ljudej* ***očen′ značim*** *dlja blagopolučija gosudarstva.*
labor these people very mean:PRESP/ADJ for prosperity country
‘These people’s labor is very significant for the prosperity of the country.’

The observations of these examples indicate that, among the participial forms, most word forms used with *očen′* can be adjectivized easily. There is an extension in meaning in the participial word forms: In addition to being qualitative, they become figurative. This is likely to weaken the semantic relatedness between the finite verbs and participles, and to lead to more adjectivized word forms. This may also imply that a qualitative meaning of an adjectivized participle is more measurable than is an initial meaning of the base verb.

The analysis of the contexts for the constructions in Table 4.7 shows that ambiguous participles occurred in both the constructions with more frequent finite verbal and participial word forms. At the same time, there were generally no occurrences of the unambiguous participles used with *očen′* in the constructions with more frequent participial word forms. Only finite verbal forms and adjectivized word forms derived from the verbs *znat′*, *zapominat′sja*, *zaxvatyvat′*, and *značit′* were used with *očen′*.

The ambiguous participles that appeared to be adjectivized had no complements and/or adjuncts, and had undergone a shift in lexical meaning. In these cases, *očen′* intensifies them as adjectives, and not as participles that have specific semantics. In the examples shown above, the unambiguous participles have complements and/or adjuncts, while the ambiguous participles have neither. These observations may imply that *očen′* is not the most salient factor of adjectivization, as it intensifies both the verbal and adjectival meanings that can be measured. At the same time, in contexts that show no other formal signs of verbal behavior (complements and adjuncts), *očen′* tends to indicate that a participle is adjectivized. Finally, the more frequent use of participles with *očen′* implies that *očen′* is attracted more to a gradable qualitative meaning in adjectivized participles than to a gradable verbal meaning in finite verbs.

2.3 Discussion and summary

The analysis of the adverbial *ocen′* constructions confirms that present finite verbal forms are used with this adverb. Their corresponding present participial word forms also combine with *ocen′*, although less frequently. The *očen′* + PRESP and *očen′* + PRS V constructions are scarce across the Araneum corpus compared to the other uses of present verbal and participial word forms. The compatibility of both types of constructions is enabled by the gradable semantic component; that is, the semantic classes (for example, the psychological, mental, and perception domains) of the base verbs.

On closer inspection, *očen′* + PRESP and *očen′* + PRS V were found to differ significantly in their ratios. This may indicate that it may not be as typical for a participle to combine with an intensifier as it would be for a verbal word form. There is a strong association between the ratio of the constructions and the semantic classes of the base verbs, although the association is more significant for the *očen′* + PRS V than for the *očen′* + PRESP constructions. It follows that high- and low-frequency uses are affected by the type of semantic class. The analysis of the box plots for the distribution of each construction showed that the verbs of the psychological domain accounted for more frequent uses of both *očen′* + PRS V and *očen′* + PRESP constructions. The verbs of change of state explain the more frequent uses of the *očen′* + PRESP construction, while the verbs in the mental domain do so for the *očen′* + PRS V construction; see Figures 4.4 and 4.5.

These findings support the assumption of the morphosemantic approach, which states that the semantics of the base verbs may lead to adjectivization reflected in the syntactic context. In this case, the class of **psychological**, **mental domains**, and **change of state** may enable the compatibility of the participles with *očen′*. The differences between the ratios of *očen′* + PRESP and *očen′* + PRS V may also imply that the kinship between the semantic classes of the base verbs and participles in the *očen′* constructions is weakened.

The qualitative analysis shows that more frequent *očen′* + PRESP constructions only consisted of adjectivized participles (for example, *zaxvatyvajušicij* ‘exciting, captivating’ and *značimyj* ‘meaningful’), while less frequent constructions can include both unambiguous and adjectivized participles. The ambiguity is manifested in the lack of complements and/or adjuncts and extension of lexical meaning. Extension seems to be an important factor, as it differentiates between (a) unambiguous participles that combine with *očen′* in the same way as their finite verbal forms, and (b) ambiguous participles with an extended meaning, which combine with *očen′* as adjectives. Thus, the gradable semantic components (psychological and mental domains) in the base verbs may be extended in adjectivized participles, and make it easier for them to combine with *očen′*.

Overall, the lexical semantics and the extension of lexical meaning (see the morphosemantic approach) appeared to be crucial factors for adjectivization in this corpus study. Furthermore, the status of *očen′* is conditional and does not necessarily indicate that a lexeme is adjectivized.

Its use indicates whether a participial word form has a gradable semantic component and/or its meaning is extended.

Očen′ is one of the adverbs of measure and degree that are used to intensify adjectival or verbal meanings. Adverbs such as *absoljutno* ‘absolutely’, *sovsem* ‘completely’ and krajne ‘extremely’ are also included in this class, and typically combine with adjectives. This is why I used them and other adverbs of the same class as a constraint in the Russian CG.[24]

3 Morphosyntactic properties and corpus frequency

This section explores the relationship between the ambiguity of participles and their morphosyntactic properties, as reflected by corpus frequency. The morphosyntactic properties include tense, voice, transitivity, and aspect. They were studied based on the interaction of two frequency distributions, namely the rank-frequency distribution of verbal lemmas and the ratio of participial word forms to the finite and infinitival word forms of their base verbs. The rank-frequency distribution consisted of verbal lemmas that were ranked according to decreasing frequency; that is, from high-frequent to low-frequent verbal lemmas. The set of participles and verbs was extracted from the disambiguated version of the RNC.

In this analysis, I examine whether the morphosyntactic properties of participles and their frequency in comparison to the other verbal forms in the corpus accounted for their ambiguity. In terms of pervasiveness, I investigate how many participles could be derived from their corresponding base verbs, expressed by the ratio of participles to their corresponding base verbs. Finally, it was deemed essential to determine whether there is a relationship between the frequency of the base verbs and the ambiguity of participles. These questions were addressed in order to understand how the corpus frequencies of participles characterize the ambiguity and pervasiveness of participles by conducting a quantitative analysis of the participles based on the examples taken from the RNC. If the word-frequency distribution is associated with POSs, the corpus frequencies of word forms and lemmas can enable a distinction of the POSs they represent.

The primary interest in morphosyntactic properties is in the positive/negative effect of the morphosyntactic properties on adjectivization.[25] Depending on its morphosyntactic properties, a participle may favor or disfavor adjectivization. For example, past passive and present active participles are prevalent among adjectivized participles due to the interaction of the tense and voice features. Past passives tend to be adjectivized more often because the perfective aspect and the passive voice reinforce the resultative meaning of the past tense, while the passive voice alone is rarely used with agents. In addition, past passive perfective verbs also have a figurative

[24] For example, the constraint is used in 16 out of 30 CG rules for removing participial readings and its effect proved to be beneficial when testing the rules on the corpora prior to the disambiguation, see Chapter 5, Section 4.2.5.1 for more details.

[25] See Chapter 3, Section 4.2.2.

meaning (Kalakuckaja, 1971), which is a great attractor of adjectivization. As a comparison, past active participles favor adjectivization less due to the combination of the past tense and the imperfective aspect that, together, intensify the processual temporal meaning, and the active voice, which is not associated with the reduced argument structure and lack of verb government.[26] The categories of transitivity and aspect have a secondary effect on adjectivization compared to tense and aspect.[27] Transitivity affects the argument structure of a participle, which manifests itself in the syntactic context in which it is used. The intransitive use alone indicates the reduction of the argument structure: As participles formed from intransitive verbs do not have slots for objects, they do not have complements, and are likely to be adjectivized according to both syntactic and morphosemantic approaches. By contrast, participles formed from transitive verbs are expected to be inclined heavily towards verb government, and thus evade adjectivization.

The pervasiveness of participles and the frequency rank of the base verbs do not seem to have been discussed previously in the literature on adjectivization. For example, one may speculate that (a) a high rank of finite verbal word forms implies a high number of their corresponding participial word forms, and (b) there must be more ambiguous word forms among high-frequency participles than among mid or low-frequency ones. Both claims are based on Zipf's law of systematic frequency distribution and the word-frequency effect[28] (Zipf, 2016, 1999). The law assumes that the frequency of occurrence of a word is almost an inverse power law function of its rank.[29] Piantadosi (2014) showed that the frequency of the POS tags from the Penn Treebank appeared to match the 'near-Zipfian law'[30] well. The observed patterns suggest that the word-frequency distribution could have an impact on some of the general mechanisms behind syntactic categories. A more specific claim (b) refers to a meaning-frequency law (Zipf, 1945) that illustrates the tendency of more frequent words to be polysemous, and universal across languages (supported by Ferrer-i-Cancho 2016; Casas et al. 2019). This complies with the idea that a diversity of expressions is facilitated while simplicity of use is preserved (van Rooij et al., 2013).

The assumptions in the research questions were evaluated based on corpus distributions and via statistical analysis. The distribution of the verbal word forms and their corresponding participial word forms allows one to see how frequently verbs form participles, and how the pervasiveness of participles is related to adjectivization. The distribution of transitivity, voice, and tense features across participles may also highlight distinct properties of participles with regard to the base verbs.

Section 3.1 describes the sources of the corpus data, and the preparation and organization

[26] The perfective suffix *-nu-* also resists adjectivization and those past active participles derived from imperfective verbs are difficult to adjectivize (except for *byvšij* 'former-PP/ADJ' synonymous with *prošlyj* 'last-ADJ' and *staryj* 'old-ADJ').

[27] See Chapter 3, Section 4.2.2.

[28] The word-frequency effect implies that more frequent words are processed faster than are less frequent ones.

[29] Zipf's law asserts the relationship between the frequency of a word and its rank: The most frequent word ($r = 1$) has a frequency proportional to 1, the second most frequent word ($r = 2$) has a frequency proportional to $1/2$, the third most frequent word has a frequency proportional to $1/3$, and so forth (Piantadosi, 2014).

[30] Piantadosi (2014) uses the term to mean frequency distributions for which this law holds, at least approximately.

thereof to obtain distributions. Section 3.2 presents a qualitative analysis of the frequency distribution of participial and finite verbal word forms. It focuses on the high-frequency, mid-frequency, and low-frequency ranges. Section 3.3 discusses the statistical analysis based on the frequency features of participles and verbs, and the morphosyntactic properties of participles.

3.1 Data and sources

I defined two types of distributions in order to test the hypotheses and answer the research questions. The first type is frequency distribution, based on the normalized corpus frequencies of verbal lemmas and their corresponding participial word forms. I grouped finite and infinitival verbal forms, and their corresponding participial word forms according to the verbal lemmas. These grouped word forms will be referred to henceforth as participial lemmas and verbal lemmas. The second type corresponds to the distributions based on the morphosyntactic properties of participles (transitivity, voice, tense, and aspect).

For each distribution, I studied a small set of participles in three to six representative examples drawn from the corpus. To understand the extension of lexical meaning and adjectivization in more detail, I analyzed these participles by taking their syntactic contexts into consideration, as these contexts clearly show whether participles (a) are used figuratively, (b) are adjectivized, or (c) both. For case (a), I manually investigated whether there was any shift in the lexical meaning of participles in some of the examples within these distributions. In this sense, syntactic context functions as a background to the morphosyntactic and semantic properties of unambiguous, adjectivized, and idiomatized participles. The analysis of these examples was qualitative, which means that conclusions about the semantics and the additional syntactic properties of the participles under discussion were based on these individual cases.

Morphosyntactic properties are expressed by the morphemes of a lemma or a lemma as a whole lexeme (in the case of syntactic function related to transitivity). These properties include inflection and POS. For example, transitivity, voice, and tense are properties that are only attributed to verbs, while case, gender, and number are properties that are shared by adjectives and verbs. I began my analysis by observing the distributions of the morphosyntactic properties across the ranking of verbal lemmas expressed in instance per million words (ipm) frequency.[31] I then ran a statistical analysis on the interaction of these properties with the frequency of verbal lemmas and the ambiguity of participles.

The datasets for the distributions were based on two types of sources:

- A list of participial suffixes (full and short forms) taken from the Russian grammar (Belousov et al., 1989) was used to retrieve participial word forms for the frequency distribution.[32]
- The weighted morphological analyzer is based on the lexicon in Zaliznjak's (2003) Grammatical Dictionary of Russian. It takes morphological readings from the dictionary

[31] It is relative frequency that is given per million words.

[32] See Appendix F.

and assigns all available readings to a given word form, including ambiguous readings ADJ and PTCP.[33]

I applied this material to retrieve verbal and participial word forms; the list of participial suffixes was used to retrieve forms, and the analyzer to annotate them. The common data were the corpus files of the gold-standard corpus of the RNC,[34] which is manually disambiguated and contains 1,190,540 word forms. Its content includes blogs from 2013, fiction, and public, science, and speech texts.

For the distributions of verbal and participial lemmas, I extracted the occurrences of ambiguous/unambiguous participles and verbal lemmas from the RNC. The occurrences of words with participial suffixes[35] in the corpus were extracted and annotated using the morphological analyzer. The annotated word forms were lemmatized and grouped as participial lemmas (verbal lemmas for ambiguous and unambiguous participles; that is, with more than one reading assigned by the analyzer) and verbal lemmas (finite verbal forms and infinitives only). The pairs of verbal lemmas and their corresponding participial lemmas were ordered according to the frequency of the verbal lemmas, and were ranked from the most to the least frequent ones. The ranking that initially ranged from 1 to 8900 was narrowed down to 6480, as beyond that, the frequencies equaled 1 and 0 (these were cases when there were no matching verbal or participial lemmas). The frequencies of verbal and participial lemmas were transformed into ipm values.

For the distributions of the morphosyntactic properties (tense, voice, transitivity, and aspect), I extracted verbal lemmas, as well as their tense, voice, aspect, and transitivity features, from the RNC, matched them to the participial lemmas, calculated the numerical ratio (that is, the *ipm frequency of participial lemmas/ipm frequency of verbal lemmas*), and ordered them according to the increasing rank of the verbal lemmas.

3.2 Pervasiveness of participles

In this section, I explore the distributions of participial lemmas based on the frequency and rank of verbal lemmas, and on the ratio of participial to verbal lemmas. As mentioned previously, verbal lemmas include finite verbal forms and infinitives; thus, there is a clear-cut contrast between participial forms and finite/infinitival ones in the analysis of the distributions. The pervasiveness of participles compared to the other verbal forms could eventually explain the frequency patterns that ambiguous and unambiguous participles exhibit in these distributions.

The distribution of verbal and participial lemmas was ordered based on the increasing rank of verbal lemmas, and was divided into three equal ranges.[36] These ranges approximated the distributions of high-frequency, mid-frequency, and low-frequency verbal lemmas. Each range

[33] For the full description of the analyzer, see Chapter 2, Section 3.4.

[34] Obtained by using the license available at http://ruscorpora.ru/corpora-usage.html.

[35] See Appendix F, Table F.3.

[36] Dividing the distribution into ranges and intervals improves the visualization.

consisted of 24 intervals with 90 lemmas per interval, as shown in Figures 4.6, 4.7, and 4.8.[37] Intervals were used to calculate the cumulative ipm frequency for the verbal lemmas and the participial lemmas found in each range. They also allowed us to observe how the number of participial lemmas increased as the rank decreased in detail, and to tentatively explain the pervasiveness and adjectivization of the participles across the high-, medium-, and low-frequency ranges.

Table 4.8 presents the morphological types[38] of the participles derived by transitive/intransitive and perfective/imperfective verbs.

Type of verb	**Type of participle**	**# participles**
transitive imperfective (TR IPFV)	active/passive, present/past	4
transitive perfective (TR PFV)	active/passive and past	2
intransitive imperfective (INTR IPFV)	active and present/past	2
intransitive perfective (INTR PFV)	active and past	1

Table 4.8: Types of verbs and the number of the morphological types of the participles they form. Morphological types refer to the combination of transitivity and aspect features a participle can represent, for example, a transitive imperfective participle.

Hypothetically, the group TR IPFV is likely to form more morphological types of participles (including adjectivized ones), followed by TR PFV and INTR IPFV. The INTR PFV group is unlikely to form as many participles as the previous groups, and these participles are also unlikely to adjectivize. I examined this assumption in the course of analyzing the examples in the distributions.

3.2.1 High-frequency range

Figure 4.6 illustrates the distribution of the high-frequency verbal and participial lemmas (high-frequency range). The range includes 24 intervals, each containing 90 verbal lemmas (the first interval is 1–90, the second, 91–180, and so on).

[37] See the graph of the entire distribution in Appendix F, Figure F.1.

[38] These types specify both transitivity (transitive and intransitive forms) and aspect (perfective and imperfective forms).

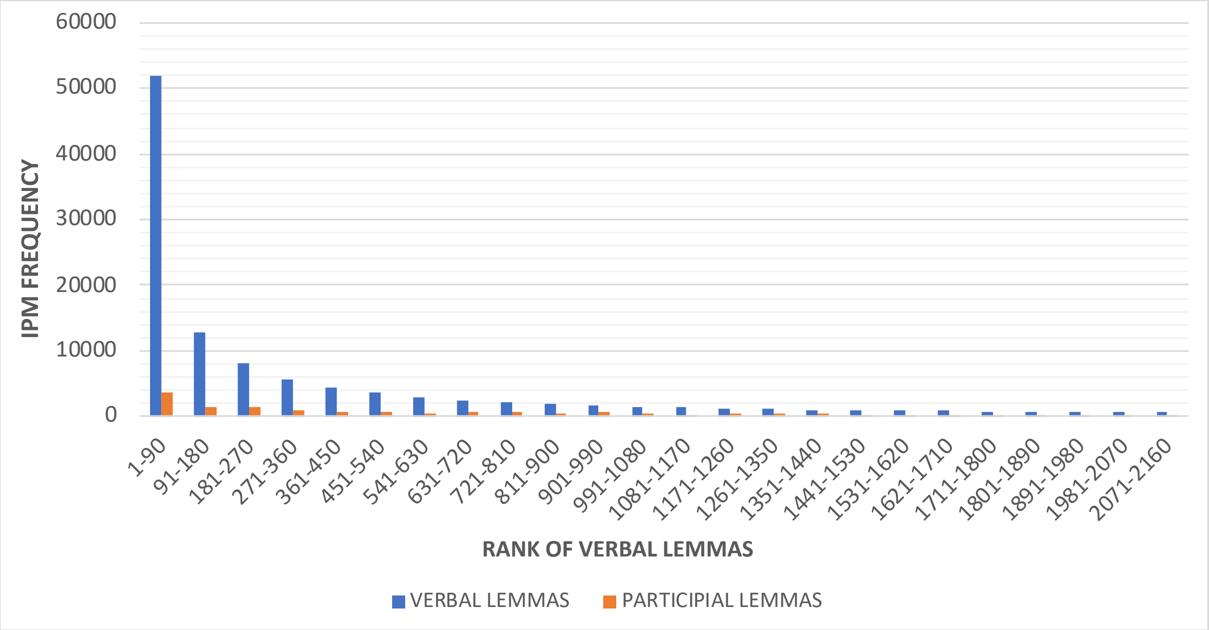


Figure 4.6: Verbal and participial lemmas ordered by the frequency rank of verbal lemmas (interval 1–2160), and the ipm frequency of verbal lemmas.

The mean of the number of the verbal lemmas (based on their ipm frequencies) is approximately 12.2 times greater than the mean of the participial lemmas across all the range. Table 4.9 indicates that 65% of all the verbal lemmas formed more verbal than participial word forms, 30.5% of the lemmas formed verbal word forms without participles, and only 4.5% had an equal number of verbs and participles or formed more participles than verbs.

Groups	**Lemmas (%)**
V > PTCP	65
PTCP = 0	30.5
PTCP ≥ V	4.5

Table 4.9: Groups of lemmas in the mid-frequency interval: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V).

Table 4.10[39] illustrates the distribution of the ten most frequent verbal lemmas with the ratios of their corresponding participial lemmas. It shows that five verbal lemmas were intransitive imperfective (INTR IPFV), three lemmas were transitive imperfective (TR IPFV), one lemma was transitive perfective (TR PFV), and one was intransitive perfective (INTR PFV).

[39] *Please Note*: I excluded the ambiguous lemmas *byt′*, *est′* ‘be, eat’ (rank 6) generated for the verbal word forms *est′* ‘eat, there is’ from the qualitative analysis in this section. The reasons were that (a) the corresponding participles for these lemmas were the same as for the lemma *byt′*, and (b) there were also frequent occurrences of *est′* as part of the set expression *to est′* ‘that is’, used as a conjunction.

Verbal lemma	**Rank**	**Ratio**	**Trans Asp**
byt′ ‘be’	1	0.014	INTR IPFV
moč′ ‘be able’	2	0.001	INTR IPFV
skazat′ ‘tell’	3	0.034	TR PFV
govorit′ ‘speak’	4	0.054	TR IPFV
znat′ ‘know’	5	0.015	TR IPFV
stat′ ‘become’	6	0.021	INTR PFV
dumat′ ‘think’	7	0.009	INTR IPFV
idti ‘go’	8	0.021	INTR IPFV
imet′ ‘have’	9	0.126	TR IPFV
delat′ ‘do’	10	0.009	INTR IPFV

Table 4.10: The top ten most frequent verbal lemmas.

The verbal lemma *byt′* ‘be’ is 71.7 times more frequent than its corresponding participial lemma, including the adjectivized active form *buduščij* ‘future’ and *byvšego* ‘former, being (in the past)’, as shown in (9).

(9) a. *rodnogo goroda, . . . byvšego teatral′noj stolicej vsego kraja*
native town, . . . be:PRESP.ACT theatrical capital entire region
‘the native town, [once] being a theatrical capital of the entire region’

b. *byvšie “gospodskie” apartamenty.*
former.NOM.PL.M manor apartments
‘former manor apartments’

In (9a), *byvšego* ‘being’ is a past active participle denoting the town as a former theatrical capital. In (9b), *byvšego* ‘former’ is an adjective characterizing the apartments that used to be manor in the past.

The verbal lemmas *moč′* ‘be able to’ had only one present active participle, *moguščego* ‘being able’, used with the verb complement *popast′* ‘fall under’, see (10).

(10) *neposedu, . . . zaprosto moguščego popast′ pod vlijanie*
restless . . . easily be.able:PRESP.ACT fall under influence
‘a restless person, capable of falling easily under influence’

Unlike most of these top ten verbal lemmas[40] with active participial lemmas only, the verb *skazat′* ‘say’ had both active and passive participial lemmas, as illustrated in (11).

(11) a. *miss Pitt Ènn, skazavšej komu-to*
Miss Pitt Ann, say:PP.ACT someone
‘Miss Pitt Ann, having said to someone’

b. *slova, skazannye nemalo raz*
words say:PP.PASS several times

[40] According to my searches of examples in the corpus.

‘the words said several times’

Some adjectivized participles in the active voice are also idiomatized, such as the active participial word form *govorjaščaja* (from *govorit′* ‘speak’) in (12). In this example, the word form *govorjaščaja* is a case of metonymy, and denotes the abstract, self-illustrative entity *familija* ‘surname’.

(12) *govorjaščaja familija*
talking.NOM.SG.F surname
‘self-explanatory surname’

Similarly, the word form *govorjaščij* ‘talking’ (from *govorit′* ‘speak’) is adjectivized and idiomatized in several other contexts given in (13).

(13) a. *govorjaščij slon* ‘talking elephant’
b. *govorjaščie koški* ‘talking cats’
c. *“govorjaščij” perečen′* ‘representative list’
d. *govorjaščaja familija* ‘self-explanatory surname’

The examples in (12) and (13) show ambiguous cases of the present active participial word forms with an adjectival meaning. In (13a) and (13b), *govorjaščij* and *govorjaščie* denotes an elephant and cats that are able to express themselves in a manner comprehensible to humans. The meaning of these word forms still relates to the meaning of *govorit′* ‘speak’. In (13c) and (13d), the word forms *govorjaščij* and *govorjaščaja* are idiomatized, as they now mean ‘self-explanatory’ and denote traits already available in the list or in the surname. The meaning of these word forms is extended from the meaning of *govorit′* ‘speak’ to ‘speak for itself’.

The present active participle *znajuščix* ‘knowledgeable’ (from *znat′* ‘know’) in *mnogo znajuščix ljudej v diplomatii* ‘many knowledgeable people in diplomacy’ conveys the state of knowing a lot, or being knowledgeable. In the other five examples, the present and past (*znajuščix*) active participles are unambiguous, as in *masterom, prevosxodno znavšim [. . .]* ‘the master perfectly knowing [. . .]’.

Some present active participles derived from the verb *dumat′* ‘think’ can also be adjectivized; see (14).

(14) a. *dumajuščij molodoj mužčina* ‘thoughtful young man’
b. *dlja umnyx i dumajuščix ljudej* ‘for smart and thoughtful people’
c. *“dumajuščej odeždy”* ‘intelligent clothes’

In (14a) and (14b), the participial word forms *dumajuščij* and *dumajuščix* ‘thoughtful’ are not idiomatized, as they still convey the meaning of ‘thinking’ related to the meaning of the base verb. In this case, the word form characterizes the person as thoughtful, or as someone who likes to think. In (14c), *dumajuščej* is idiomatized, as it is now combined with the inanimate noun *odeždy* 'clothes', and conveys the meaning of 'intelligent clothes', translated from 'clothes that do the thinking'.[41] This use also appears to be metaphorical.

I did not find adjectivized participles derived from the verbs *stat′* 'become', *imet′* 'have', and *delat′* 'do'. Some participial word forms (derived from *idti* 'go'), which formally demonstrate verbal behavior, appeared inside fixed expressions,[42] as in (15). The word forms *iduščaja* and *iduščix* 'going' are used in the fixed expressions *vperedi iduščaja nauka* 'the pioneer science' in (15a) and *daleko iduščix vyvodov* 'far-reaching conclusions' in (15b). Even though the participles in (15) are modified by the adverbs *vperedi* 'ahead' and *daleko* 'far', their use in the fixed expressions may point at the more adjectival nature of the lexical meaning of the participles.

(15) a. *vperedi iduščaja nauka*
ahead go:PRESP science
'the pioneer science'

b. *daleko iduščix vyvodov*
far go:PRESP conclusions
'far-reaching conclusions'

In the high-frequency range, there were seven verbal lemmas within the interval of 1–100 (that is, 7% of all the lemmas in this interval), and no participles.[43] The low number of verbal lemmas that had no respective participles indicates that the verbs at the top of the frequency rank tended to form participles repeatedly, with only a few cases of verbs that lacked participles.

The analysis of the distribution shows that finite verbal forms were much more pervasive than were participles in the high-frequency range. This may be explained by the fact that a participle is an uncommon type of a verbal form, as it is a hybrid and has both verbal and adjectival properties. For this reason, it always has two competitors, namely finite verbs and infinitives.

The analysis of the top ten most frequent verbal lemmas indicated that these verbs had adjectivized and idiomatized participial word forms; however, the ratios of these participles to finite/infinitive verbs were relatively low. The examples of the adjectivized participles formed from these base verbs are the present participle *buduščij*, which signifies 'future', past participle *byvšego* with the most frequent meaning 'former' or the present participle *govorjaščij* used with the extended meaning 'talking, self-explanatory'. In addition, all of the base verbs can form two or four morphological types of participles, which means that they may always have participles, albeit ones that are not used frequently.

[41] See Russian version: http://news.bbc.co.uk/hi/russian/sci/tech/newsid_1093000/1093993.stm

[42] The adverbials of place preceding participles make this expression fixed.

[43] See Appendix F, Table F.4.

3.2.2 Mid-frequency range

Figure 4.7 below illustrates the mid-frequency range of verbal lemmas and their corresponding participial lemmas.

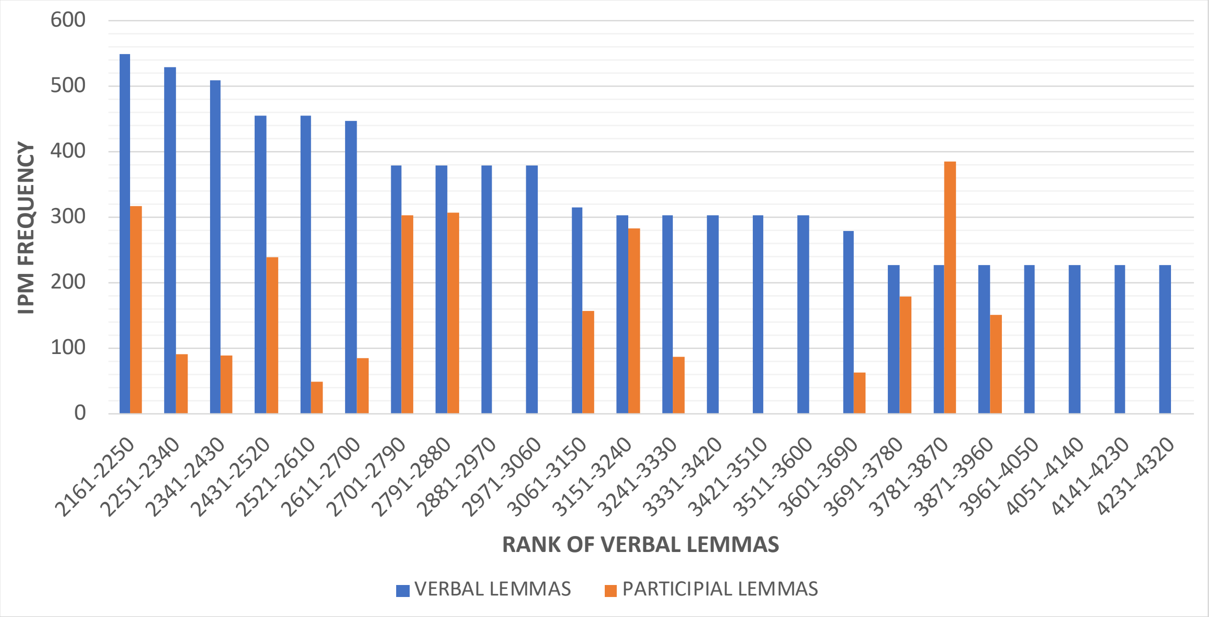


Figure 4.7: Verbal and participial lemmas ordered according to the rank of the verbal lemmas' frequency (interval 2161–4320).

The range includes 24 intervals with 90 verbal lemmas per interval, with or without participial word forms. Across all the intervals, verbal lemmas are, on average, approximately 1.1 times more frequent than are participial lemmas. Table 4.11 shows that 30.8% of all the lemmas formed more verbs than participles, 59.8% of all the lemmas had verbs only, and 9.4% had an equal number of participles and verbs, or formed more participles than verbs.

Group	Lemmas (%)
V > PTCP	30.8
PTCP = 0	59.8
PTCP ≥ V	9.4

Table 4.11: Groups of lemmas in the mid-frequency range expressed in percentages: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V).

There were both present active and passive participial forms among the lemmas in the 59.8%[44] category; thus, the group INTR PFV is representative across these intervals (based on 21 examples). Some of these word forms were also adjectivized and idiomatized, as shown in (16). For example, the present active word form *potrjasajuščix* (from the verb *potrjasat′* 'astonish') in (16a) is used as an adjective in *potrjasajuščix glazax* 'gorgeous eyes'. The word form is also a

[44] See Appendix F, Table F.5.

metonym because the initial participial meaning is extended to a qualitative one, and modifies the noun *glazax* 'eyes'. The present active word form *intrigujuščim* (from *intrigovat′* 'intrigue'), as in *intrigujuščim predisloviem* 'intriguing foreword' in (16b), is used with an abstract meaning to modify the noun *predisloviem* 'foreword'. An example of the past passive adjectivized word form is *osmyslennaja* 'sensible' (from *osmyslit′* 'apprehend'), which extends its meaning to modify the abstract noun *politika* 'policy'; see (16c).

(16) a. *potrjasajuščix glazax*
amazing.PRP.PL.M eyes
'gorgeous eyes'

b. *intrigujuščim predisloviem*
intriguing.INS.SG.NEUT foreword
'intriguing foreword'

c. *osmyslennaja politika*
sensible.NOM.SG.F policy
'sensible policy'

3.2.3 Low-frequency range

Figure 4.8 illustrates the low-frequency range with 24 intervals and 90 verbal lemmas per interval (from 4321–6480).

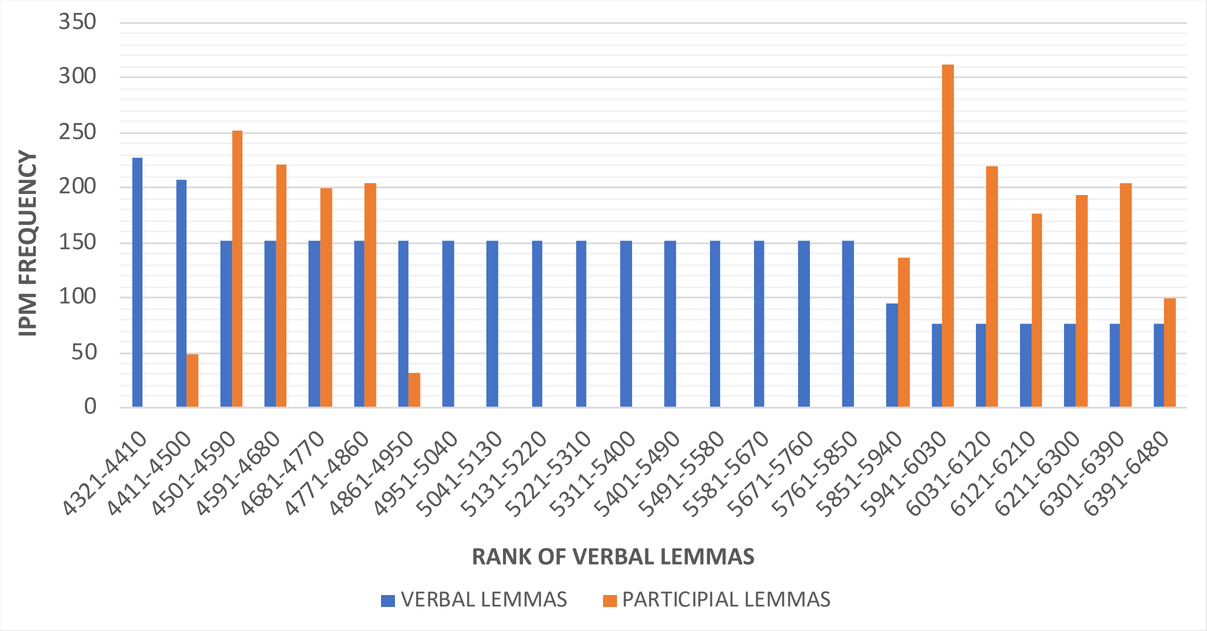


Figure 4.8: Verbal and participial lemmas ordered according to the rank of the verbal lemmas' frequency (interval 4321–6480).

Within this range, the mean number of verbal lemmas is 0.4 times greater than the number of the participial lemmas. Table 4.12 shows that about a half of all the lemmas in the range did not form participles (55.2%), less than half formed more participles than verbs (35.5%), and only 9.4% of the lemmas had fewer verbs than participles or an equal number of participles and verbs.

Group	Lemmas (%)
V > PTCP	9.4
PTCP = 0	55.2
PTCP ≥ V	35.5

Table 4.12: Groups of lemmas in the low-frequency range: more verbs than participles (V > PTCP), verbs only (PTCP = 0), and more participles than verbs or an equal number (PTCP ≥ V).

Some cases in the 55.2% of verbal lemmas without participles were *vvalit'sja* 'burst in', *soveršenstvovat'* 'improve', *utomljat'* 'tire', and *čertyxat'sja* 'swear' (see more examples in Appendix F, Table F.6). The group of intransitive perfective (INTR PFV) participles was the least pervasive across these intervals.

In the sentences with ambiguous participles, I found instances of adjectivized participles with an extended lexical meaning in the present active and past passive word forms, as illustrated in (17). The present active word form *podobajuščego* 'decent' in (17a) is an adjective that modifies the abstract noun *obrazovanija* 'education'. Although the original meaning of its base verb is abstract, it is clearly qualitative in *podobajuščego* 'proper'. The present active word form *otravljajuščix* 'poisonous' in (17b) is used as an adjective to modify *veščestv* 'substances'; its meaning is extended to a qualitative one, but it is not idiomatized. In (17c), the present active word form *podkupajuščej* 'disarming' is used metaphorically as an adjective in *podkupajuščej prjamotoj* 'disarming frankness'. The word form *podkupajuščej* developed a qualitative meaning that was further extended to a figurative one.

(17) a. *ne imejuščij podobajuščego obrazovanija*
not having proper.ACC.SG.NEUT eduction
'not having proper education'

b. *otravljajuščix veščestv*
poisonous.GEN.PL.NEUT substances
'poisonous substances'

c. *podkupajuščej prjamotoj*
disarming.INS.SG.F frankness
'disarming frankness'

3.2.4 Summary

Overall, the observations above indicate that base verbs prefer forming finite/infinitival verbal forms to participles, or only have finite/infinitival verbal forms. High-frequency base verbs form more participles per lemma, and more than a third of low-frequency base verbs form more participles than the other verbal forms. In addition, the difference between participial and verbal word forms per lemma decreases in the mid- and low-frequency ranges compared to the high-frequency one.

High-frequency base verbs (within the range of 1–100) appear to form more verbal forms than participles, although not many of the verbal lemmas completely lacked participial word forms. Most of the frequent verbal lemmas have active participial forms that can be adjectivized, with the additional extension of semantics, as in *govorjaščaja familija* 'self-explanatory surname' (metonymy) as opposed to *govorjaščij slon* 'talking elephant' (the elephant being able to talk). The total of the high-frequency base verbs appear to form mainly finite/infinitival verbal forms that are more frequent than participles (65%) or do not have any participles at all (30.5%). Eight of the top ten most frequent verbs only formed active participles.

An interesting observation is that I only identified past passive and present active participial word forms as being ambiguous in the RNC texts across all three intervals. It was generally more difficult to find instances of adjectivized participles in the mid-frequency range, which could be explained by the lowest percentage of participles being formed by the base verbs. In the low-frequency range, the difference between the frequency of verbs and participles becomes even and low. Mainly the passive participles seemed to be adjectivized by acquiring the qualitative meaning typical of adjectives. The examples of the adjectivized participles did not show any further extension of semantics in the low-frequency range.

In each frequency range, the examples of adjectivized participles in the corpus that I analyzed were mainly represented by either present active participles or past passive participles. I observed that, in the high- and mid-frequency ranges, there were more present active than past passive adjectivized participles, while in the low-frequency range, there were more past passive than present active participles. Among the verbal word forms with no corresponding participles in the mid- and low-frequency intervals, the group of intransitive perfective forms seemed to be more representative than the other groups.

3.3 Statistical analysis of the distributions

In this section, I measure the relationships of the ipm frequency of base verbs statistically: the morphosyntactic properties of their corresponding participles on one hand, and ambiguity of participles on the other. Ambiguity is modeled on the basis of the double ADJ/PTCP readings assigned by the morphological analyzer to a participial word form from the fully disambiguated RNC subcorpus. Word forms that only receive participial reading are unambiguous, and those that receive the ADJ/PTCP readings are ambiguous. This approximation of ambiguity was based on Zaliznjak's (2003) dictionary; therefore, the correspondence between participial word forms and their ambiguity reflects the occurrence of their ambiguous readings in the dictionary. Thus, for each type of distribution, there is additional information about the possible ambiguous readings for their participial word forms based on Zaliznjak's (2003) dictionary.

The relationship between frequency/morphosyntactic properties and ambiguity aimed to test the following hypotheses:

- *Null Hypothesis* H_{01}: There is no relationship between (a) the ipm frequency of base verbs and the ratio of participles, and (b) ambiguity of participles.

- *Null Hypothesis* H_{02}: There is no relationship between the morphosyntactic properties (tense, voice, aspect, and transitivity) of participles and the ambiguity thereof.
- *Hypothesis* H_1: There is a significant relationship between (a) the ipm frequency of base verbs, ratio of participles and (b) ambiguity of participles.
- *Hypothesis* H_2: There is a significant relationship between the morphosyntactic properties of participles and the ambiguity thereof.

H_2 aimed to examine the effect of tense, voice, aspect, and transitivity features. For example, ambiguity may be associated with intransitive imperfective participles that are least prevalent among the other groups, present active participles derived from high-frequency base verbs, and past passive participles derived from low-frequency base verbs (see Section 3.2 for the pervasiveness of participles).

3.3.1 Experiment design: Data and definitions of models

The statistical analysis was based on three datasets with variables that constitute the distribution discussed in Section 3. Dataset 1 consists of variables that stand for the ipm frequency of base verbs, the ratio of participial to verbal word forms (expressed in lemmas), and the ambiguity tags of participial word forms (expressed in lemmas). Dataset 2 consists of variables for the ipm frequency of base verbs, tense and voice properties of participial lemmas, and the ambiguity tags of participial lemmas. Dataset 3 contains variables for the ipm frequency of base verbs, the transitivity and aspect properties of participial lemmas, and the ambiguity tags of participial lemmas. Table 4.13 illustrates that the length of each dataset expressed in the number of rows appears to be very similar.

Datasets	**Variables**	**N**
Dataset 1	*freq*	3,336
Dataset 2	*tense.voice*	3,384
Dataset 3	*trans.asp*	2,678

Table 4.13: Size of the datasets, expressed in number of rows (N). Each dataset contains variables used in the statistical experiment. For example, tense and voice variables are grouped as *tense.voice* in Dataset 2.

In each dataset, I capped the outliers from numeric variables (ipm frequency, ratio) by replacing the values that were outside of the lower limit with the value of the 5th percentile, and those that were above the upper limit with the value of the 95th percentile. I also logarithmically transformed the ipm frequency scores in order to make the distribution similar to normal. Some missing entries[45] in Dataset 3 for ambiguous/non-ambiguous scores were replaced with zeros

[45] Missing entries resulted from the annotations of the analyzer in which no participial or ambiguous participial/adjectival readings were identified.

(non-ambiguous).

Table 4.14 illustrates the mean scores and counts in Datasets 1, 2, and 3 after the removal of outliers, logarithmic transformation, and the replacement of missing entries.

Table 4.14: Estimate of variables in the datasets

ipmVerb	**ratioPtcp**	**ambiguity**	
		0	**1**
M: 2.046	*M*: 0.773	2942	394

(a) Dataset 1

ipmVerb	**tense**		**voice**		**ambiguity**	
	praes	**praet**	**act**	**pass**	**0**	**1**
M: 1.745	1212	2172	1909	1475	2918	466

(b) Dataset 2

ipmVerb	**transitivity**		**aspect**		**ambiguity**	
	tran	**intr**	**pf**	**ipf**	**0**	**1**
M: 2.039	2001	677	1375	1303	2304	374

(c) Dataset 3

Note. *M*: mean score, *ipmVerb*: ipm frequency of verbal lemmas, *ratioPtcp*: ipm-based ratio of participial to verbal lemmas, *praes*: present tense, *praet*: past tense, *act*: active voice, *pass*: passive voice, *tran*: transitive use, *intran*: intransitive use, *pf*: perfective aspect, *ipf*: imperfective aspect, *0*: unambiguous reading, *1*: ambiguous reading.

The mean values were assigned to the continuous variables (*ipmVerb* and *ratioPtcp*), and the counts to the variables with binary and categorical values (*ambiguity*, *tense*, *voice*, *transitivity*, and *aspect*). The mean value for *ipmVerb* was similar across all the datasets; there were more unambiguous than ambiguous tags, more past than present word forms, more active than passive word forms, more transitive than intransitive, and more perfective than imperfective word forms. Note that I only estimated the counts for variables for each dataset, and did not calculate the overlaps among them. In addition, the *ipmVerb* scores were taken from the same source and were shared across all the datasets.

Table 4.15 shows the variables that were used to analyze the distributions and in the statistical models. Ambiguity is expressed via the dependent binary variable *ambiguity* (1 (ambiguous) and 0 (unambiguous)). The variables *ratio* and *ipm frequency* are continuous (that is, variables with numeric values). The variables *tense*, *voice*, *transitivity*, and *aspect* are categorical, and consist of two levels; for example, the variable *tense* has the levels *praes* (present) and *praet* (past).

Categorical variables	**Levels**	
	0	**1**
tense	*praes*	*praet*
voice	*act*	*pass*
transitivity	*intr*	*tran*
aspect	*ipf*	*pf*
Continuous variables		
ipmVerb		
ratioPtcp		

Table 4.15: Description of the variables.

3.3.2 Overview of the distributions

In this section, I outline and investigate the distributions of the categorical variables. The overview allows us to have a first impression of how ambiguity is associated with morphosyntactic properties in the interaction with the ratio of participles and the ipm frequency of the verbs.

Figure 4.9 illustrates the relationship between ambiguous and unambiguous participial lemmas (or participles), the log-transformed ipm frequency of their corresponding verbal lemmas (or base verbs), and the ratios of participles. The scatter plot models the potential ambiguity of participles that is not distributed randomly. It shows that most of the unambiguous participles cluster around the higher ipm frequency (*ipmVerb* with scores of between 3 and 5) and a low ratio (with scores of between 0 and 1), while most of the ambiguous participles are scattered in higher ratio (*ratioPtcp* > 0.5) and higher ipm frequency (*ipmVerb* > 4.5). The number of ambiguous participles decreases between the scores 3 and 4, and begins to cluster around the lower ratio. This may have two implications: (a) that low-frequency verbs form more ambiguous than unambiguous participles, and (b) the pervasive participles that considerably surpass their respective finite/infinitival forms in number per verbal lemma tend to be adjectivized. In brief, low-frequency verbs tend to have a larger proportion of participles that also prefer to be adjectivized. While I did not continue to investigate the causal relationships conveyed by these implications, they do stress the importance of frequency in explaining adjectivization.

The base verbs with the highest frequency also formed many participles that could be adjectivized, although the ratio of these participles was lower than was the ratio among the base verbs with lower ipm frequency.

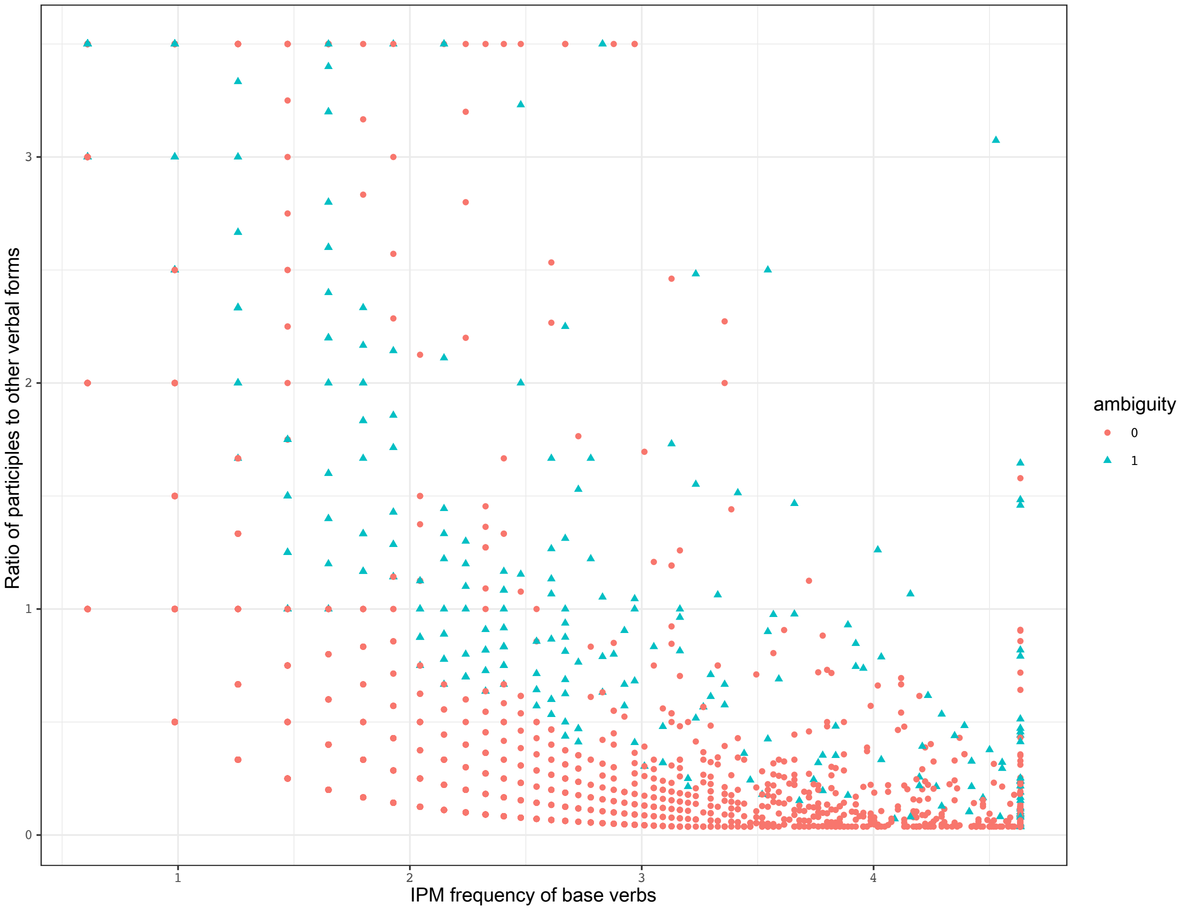


Figure 4.9: The distribution of ambiguous and unambiguous participial lemmas across the ratio of participial lemmas to verbal lemmas, and the ipm frequency of verbal lemmas.

Figure 4.10 illustrates the distribution of lemmas in the interval between 4 and 4.7 (high-frequency range). The overall number of unambiguous participles was higher than was the number of ambiguous ones, although the ratio of unambiguous participles to the finite/infinitival forms was apparently lower than was the ratio of the ambiguous ones. The base verbs with the highest ipm frequency (> 4.6) formed more participles per verbal lemma compared to the verbs with lower frequencies. There were more ambiguous than unambiguous forms among these participles. Thus, the base verbs in the highest frequency rank formed a large number of ambiguous participles.

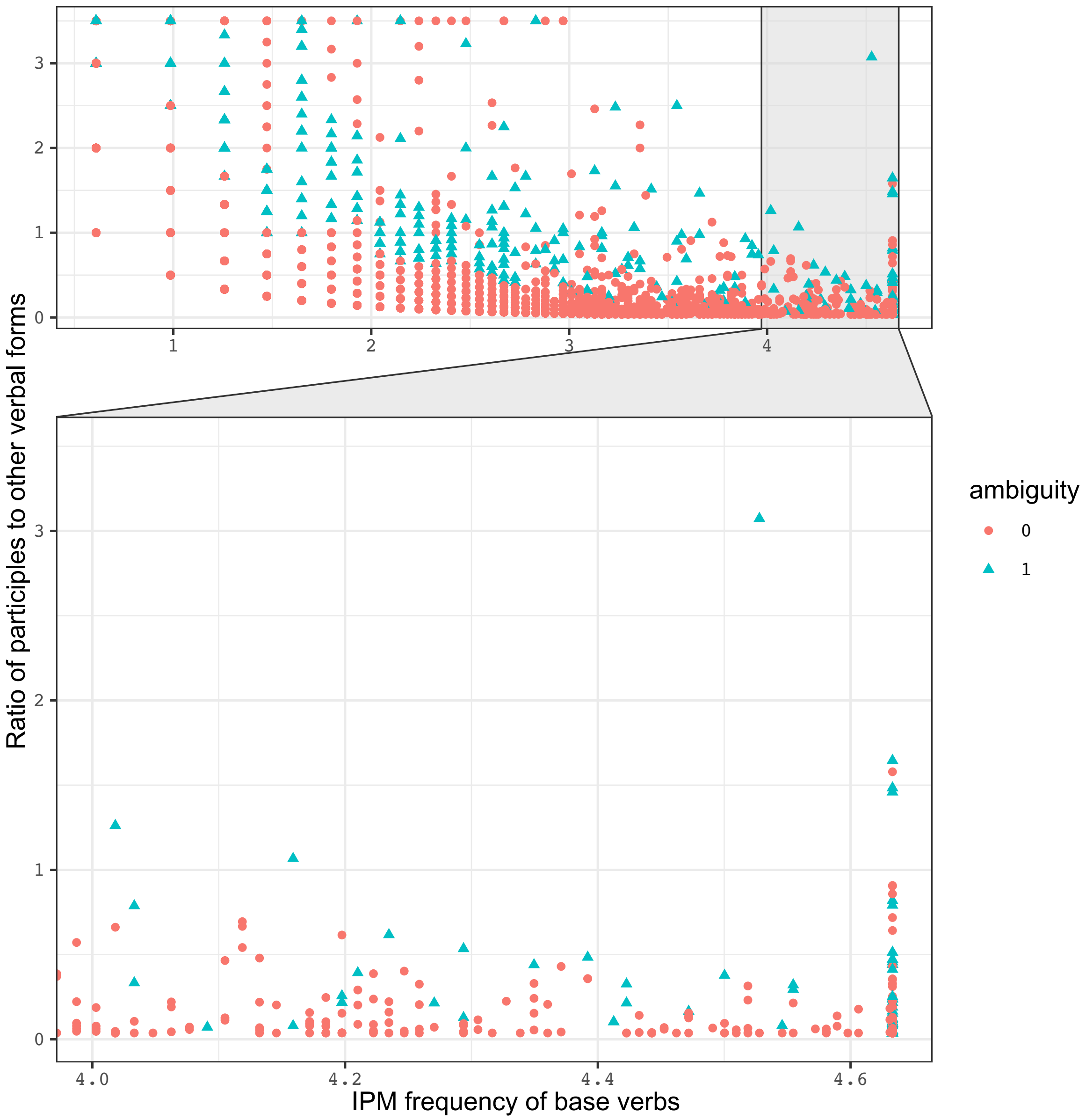


Figure 4.10: Ipm frequency interval of 4–4.7.

Figures 4.11 and 4.12 display the distributions of two groups of participles that interacted with the ratio and ambiguity features (placed on the facets 0 (unambiguous) and 1 (ambiguous)). The first group was defined on the basis of *tense* and *voice* variables to represent the morphological forms of participles such as present active, past passive, present active, and so on. The second group was based on *aspect* and *transitivity*, as these two features also combine to represent the participles that can have one, two, or four forms based on tense and voice (see Table 4.8). For example, intransitive perfective participles can only be active and past. In Figure 4.11, the participles are grouped by tense and voice (PST PASS, PST ACT, PRS PASS, PRS ACT), and in Figure 4.12 by transitivity and aspect (TR PFV, TR IPFV, INTR PFV, INTR IPFV). The distributions are represented by the violin box plots that consist of box plots and density curves.[46]

[46] The density curve is plotted as an approximation of the probability density function, which is computed using a

The white areas and black bars indicate the inter-quartile range and the median value, respectively. The mean values for each morphological type of participles are presented in red. The red and blue curves show the distribution shape of the variables.

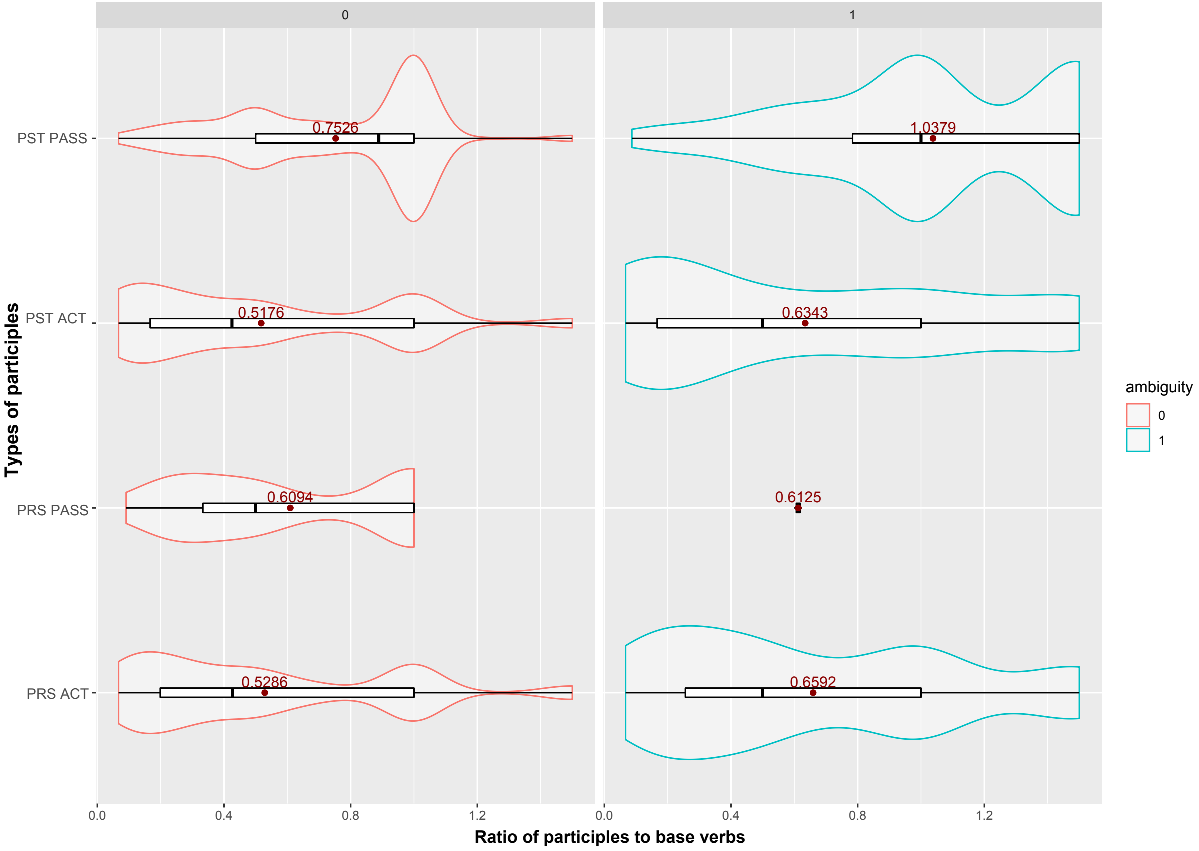


Figure 4.11: Violin plots displaying the distribution of the ratios of participles grouped according to ambiguity (0 and 1), tense and voice, and ordered by their ratio to finite/infinitive verbs.

In Figure 4.11, it can be seen that the ranges of the ratios for unambiguous and ambiguous participles overlap. This means that there are no considerable differences between the means of unambiguous and ambiguous participles. Most of the distributions are right-skewed, except for the distribution of unambiguous PST PASS participles, which is left-skewed. The largest difference is in the shape of the distribution of unambiguous and ambiguous PRS PASS participles. The sample size of ambiguous PRS PASS participles must have been too small to be estimated in terms of density. Overall, the mean value was higher for ambiguous participles than for unambiguous ones. For example, the mean was the highest for PST PASS participles (*M* = 1.0379), and was second highest for PRS ACT (*M* = 0.6592). In addition, the density curves for the ambiguous participles (except for PRS PASS) are visibly much wider than are those of the unambiguous ones. This means that the frequency of these participles was greater than was the frequency of unambiguous participles in the wider parts of the density curves.

kernel density estimation (KDE).

Figure 4.12 similarly presents the distribution of unambiguous and ambiguous participles grouped according to transitivity and aspect. In this case, the distributions of the unambiguous and ambiguous participles differed more among their groups (for example, TR PFV and TR IPFV) compared to the distributions in Figure 4.11. For example, the 95% confidence interval is shorter for unambiguous than it is for ambiguous INTR PFV participles. Within the confidence interval, unambiguous INTR PFV participles cluster around lower ratios (between the ratios of 0 and 2), while the ambiguous ones cluster around a higher ratio (between the ratios of 0 and 4). The ambiguous distributions on the right-hand side have much wider density curves than do those on the left-hand side, although all of them seem to be right-skewed to a greater or lesser degree. The mean ratios for ambiguous participles were greater than those for ambiguous ones (for example, *M* = 1.07842 versus *M* = 1.95299 for TR PFV). The ambiguous TR PFV participles had the highest mean ratio, followed by the ambiguous INTR PFV and TR IPFV participles (*M* = 1.36998 and *M* = 1.18229). The lowest mean ratio was found for the unambiguous INTR IPFV participles (*M* = 0.55769).

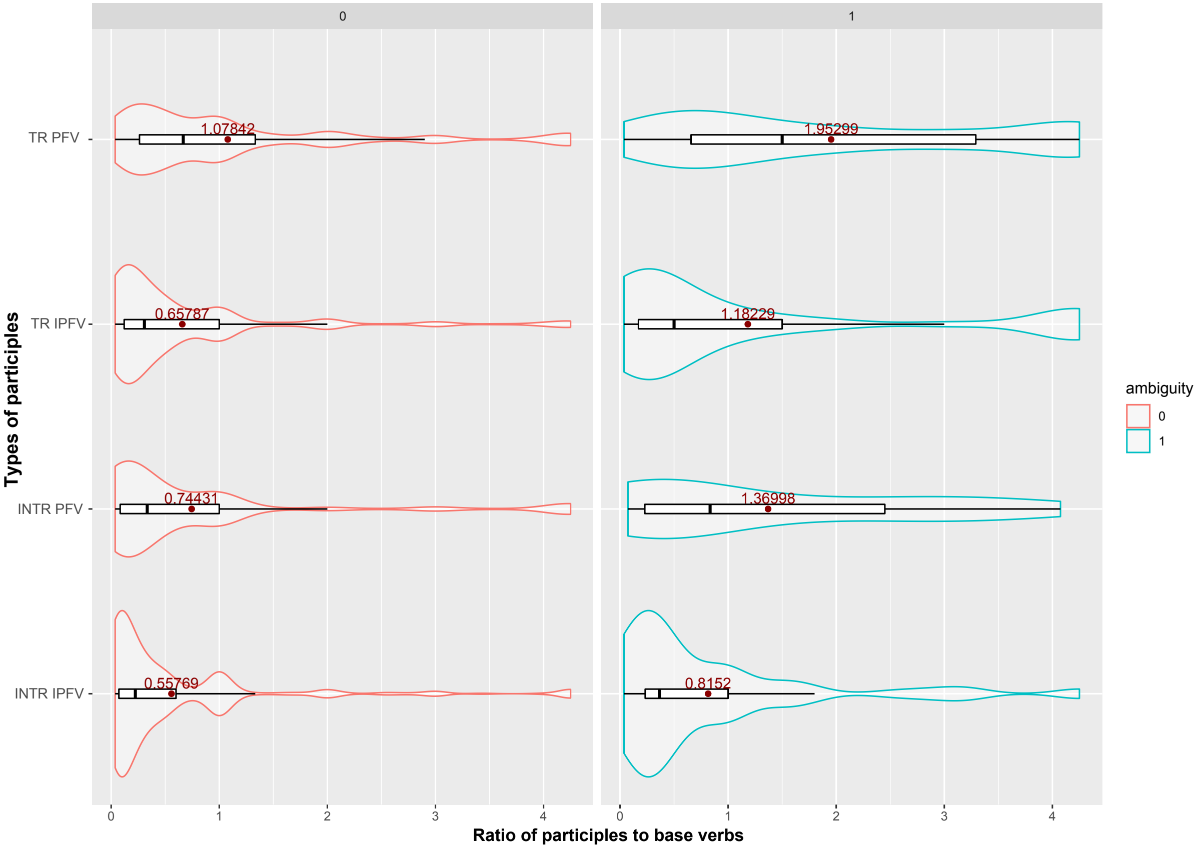


Figure 4.12: Violin plots displaying the distribution of the ratios of participles grouped according to ambiguity (0 and 1), transitivity and aspect, and ordered by their ratio to finite/infinitive verbs.

The observations in Figures 4.11 and 4.12 indicate that ambiguous forms were generally more pervasive in the corpus than were unambiguous ones; the ambiguous past passive, followed by the present active participles, were the most frequent compared to the other distributions. This concurs with the assumptions of the morphosemantic approach in which past passive and present active adjectivized participles are considered to be the most numerous group (Kustova, 2012). The passive voice reduces argument structure and the present tense leads to the development of an atemporal generic meaning; thus, both of these factors favor adjectivization (Kalakuckaja, 1971). The ambiguous transitive and intransitive perfective participles had the highest ratio compared to the other groups of participles. Some aspects relating to the morphological types of participles and their internal properties that affect adjectivization may partly explain the observations. First, transitive perfective verbs can have two morphological types of participles,[47] which implies that they are likely to produce an average number of participles. Second, the intransitive use, which accounts for the reduced argument structure, and the perfective aspect, which conveys resultative meaning, concur in adjectivization. However, the highest mean ratio of TR PFV can probably be related to the low rank of the base verbs rather than being explained by the grammatical properties of these features.

The observations of groups of participles can be linked to the general distribution of ambiguous/unambiguous forms in Figure 4.9. Past passive participles mainly corresponded to low-frequency verbs, while present active participles corresponded to mid- and high-frequency verbs. Almost all the unambiguous participles (except for TR PFV) grouped according to transitivity and aspect had a ratio of less than 1, which corresponded to mid- and high-frequency base verbs, while the ambiguous ones with a ratio of greater than 1 tended to cluster around low- and mid-frequency base verbs. Thus, based on the observations above, the tendency appears to be that a higher ratio of ambiguous participial forms (grouped according to the morphosyntactic properties) is associated with a lower rank of the base verb, and a lower ratio of ambiguous participles is associated with the higher ranking verbs. An exception is a bigger cluster of participial forms around the most frequent base verbs in Figure 4.10. This observation implies that the most frequent verbs formed more participles than did the rest of the high-frequency verbs, and that a great number of these participles were ambiguous.

3.3.3 Statistical modeling

I defined three statistical models based on these datasets, with one model for each dataset. The model definition meets the objectives of the alternative hypotheses; that is, to identify whether there was a significant relationship between (a) the ipm frequency of the base verbs and the ambiguity of participles (H_1), and (b) the morphosyntactic properties[48] and the ambiguity of participles (H_2). In other words, the models aimed to predict the effects of the frequency and morphosyntactic factors on ambiguity.

- **Model 1 (frequency, ratio)**: Ambiguity as a dependent variable explained by verb frequency and the ratio of participles to verbs, as well as the interaction of the ratio and ipm frequency.

[47] Transitive perfective verbs form past active/passive participles, while intransitive perfective verbs form one morphological type, that is, past active participles.

[48] These properties represent morphological types of participles, among others.

Model 1 with the interaction term is as follows:
*ambiguity = ratio + ipm frequency + ratio*ipm frequency*

- **Model 2 (tense, voice, frequency)**: Ambiguity, as a dependent variable, is explained by the tense, voice, ipm frequency, and interactions of (a) tense and voice, (b) tense and ipm frequency, and (c) voice and ipm frequency. Model 2 with the interaction terms is as follows:
  *ambiguity = tense + voice + tense*ipm frequency + voice*ipm frequency + tense*voice*
- **Model 3 (transitivity, aspect, frequency)**: Ambiguity, as a dependent variable, is explained by the transitivity, aspect, and interactions of (a) transitivity and ipm frequency, (b) aspect and ipm frequency, and (c) aspect and transitivity as independent variables. Model 3 with the interaction term is as follows:
  *ambiguity = transitivity + aspect + transitivity*ipm frequency + aspect*ipm frequency + transitivity*aspect*

In Model 1, the interaction of the ipm frequency and the ratio allows us to check whether the ratio of participles to high-frequency, medium-frequency, and low-frequency verbs can have an effect on the ambiguity of participles.[49] In Model 2, the interaction shows how tense and voice properties associated with the morphological types of participles can affect ambiguity. As discussed in Chapter 3, present active and past passive participles represent the most pervasive group of adjectivized participles; therefore, I expect the effect of the past/present tense dependent on the active/passive voice to affect the ambiguity. The interaction of these features with the ipm frequency of the base verbs is related to the observations made previously in Section 1.3: Present active participles derived from high-frequency verbs and past passive participles derived from low-frequency verbs tended to be used as ambiguous word forms in the context of the RNC. In Model 3, the variables of transitivity and aspect were chosen deliberately to reflect their effect on adjectivization[50] and the pervasiveness of participles in terms of their morphological types (see Table 4.8). Transitivity alone affects the capacity of the argument structure, while aspect favors/disfavors adjectivization in combination with transitivity. The joint features of transitivity and aspect account for the pervasiveness of the morphological types, together with the resultativity/temporality of their grammatical meanings. Nonetheless, the transitivity and aspect features in Model 3 may be secondary in terms of explaining ambiguity, as opposed to the morphological types of participles in Model 2.

All the models had *ambiguity* as a dependent binary variable. Frequency and morphosyntactic properties were the main effects on ambiguity, as expressed by the independent fixed variables (or predictors) *ipmVerb*, *ratioPtcp*, *tense*, *aspect*, and *voice*. Model 1 included two continuous predictors, namely *ratioPtcp* and *ipmVerb*. Models 2 and 3 included two categorical predictors, which were tense and voice in Model 2, transitivity and aspect in Model 3, and one continuous

[49] See Section 3.2 for further discussion.
[50] See Chapter 3, Section 4.2.2.

predictor (the ipm frequency of verbal lemmas).

To provide a math-based assessment of the significance of the factors of ambiguity, I resorted to a binary logistic regression.[51] A binary logistic regression addresses the continuous/categorical independent variables and the binary categorical dependent variables that do not follow a bell-shaped normal distribution (see Figure 4.13). The model is aimed at explaining a binomial outcome (X), with one or more independent variable(s) (Y) by establishing the relationship between the dependent and independent variables. More specifically, it examines the relationship of a binary outcome (for example, alive/dead, success/failure, and yes/no) with one or more predictors, which may be either categorical or continuous (Ranganathan et al., 2017).

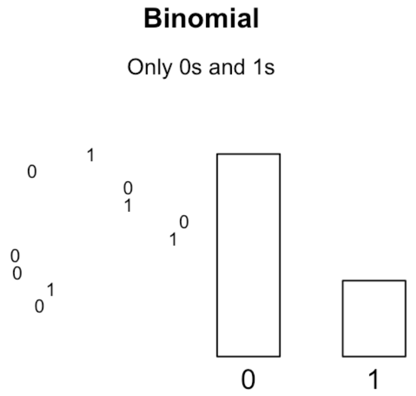


Figure 4.13: Binominal distribution in GLM.

The assumptions of the binary logistic regression model are a binary dependent variable, the independence of each data point, the correct distribution of the residuals, the correct specification of the variance structure, the linear relationship between the independent variables, and the logit[52] of the response variable.[53]

Using the function *glm()* with the option *family=binomial*(*link="logit"*), I fitted the defined models using the independent variables and their interaction in the parameters. For Models 2 and 3 with the interactions, I sum-coded[54] the categorical response variables to also take the coefficients of the predictors used in the interactions into consideration. In other words, the intercept of the models was not the mean of one level of a factor, but the mean of values of the factors in each dataset. I also scaled the values of the predictor *ipmVerb*, which allowed for better coefficients to be obtained. I used the summary function that provided the inference of the parameters of these models, namely ambiguity (ambiguous, unambiguous) as a dependent

[51] The model is available via the *glm()* function with the option (*family=binomial*(*link="logit"*)) in *R* (R Core Team, 2019).

[52] A Logit function represents probability values of 0 to 1 between negative and positive infinity.

[53] Available at: https://biologyforfun.wordpress.com/2014/04/16/checking-glm-model-assumptions-in-r/

[54] Sum coding compares the mean of the dependent variable for a given level to the overall mean of the dependent variable over all the levels. After sum coding, the intercept is the mean of means (grand mean) of the dependent variable for each level (assuming the two variables have independent effects). Available at: https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/

variable explained by the predictors. The R code for processing the datasets, setting up, and running the models and their output is presented in Appendix G.

3.3.4 *GLM* model performance

Table 4.16 below provides a summary of the inference of the parameters of the binary GLM Models 1, 2, and 3. The *parameter* column contains the names of the predictors,[55] *:* indicates the interaction of predictors, and the *P-value* column contains p-values for the significance of the predictors and their interactions. Overall, the main effects of ipm frequency and ratio, tense, and voice appeared to be strong predictors of ambiguity in Models 1 and 2, respectively. The effects of transitivity and aspect were not associated strongly with *ambiguity*, apart from interactions with the ipm frequency in Model 3. Standard error (*SE*) scores across all the models remained within the range of 0.07 and 0.048; the scores were relatively low due to the coefficient estimate being calculated imprecisely.

Model 1 estimated the positive coefficients (Estimate) for the ratio and ipm frequency, which indicates their positive association with ambiguous readings (1.141 and 0.796, respectively). This implies that, as the ratio or ipm frequency increases, it is more likely to be attributed to ambiguous participles than to unambiguous ones. The coefficient for the interacting ratio and ipm frequency *ratioPtcp:ipmVerb* was positive (0.324), but was lower than the coefficients for ratio and ipm frequency considered separately. This may imply that the interaction effect of a higher ratio and ipm frequency on ambiguity is not as strong as is the individual effect of the predictors. In Model 1, the effect of ipm frequency, the ratio of participles to base verbs, and their interaction was significant in predicting ambiguity (p < 0.001). This means that the ratio of participles to the base verbs dependent on the ipm of the base verbs also has a significant effect on ambiguity.

In Model 2, the estimated coefficient for the past tense was negative (-0.525), which indicates its strong association with unambiguous readings. The coefficient became slightly higher, but still negative, when the past tense depended on the ipm frequency (-0.019). The individual effect of the past tense (p = 0.004) and the passive voice (p = 0.019) was significant in terms of predicting ambiguity. According to the estimated coefficients, the past tense favors unambiguous readings (-0.525), while the passive voice favors ambiguous ones (0.429). Despite the fact that the interactions of past tense/passive voice with the ipm frequency had the lowest standard error scores (*SE* = 0.062 and 0.064, respectively), their p-values were large (p > 0.5). In addition, the effect of the past tense depending on the passive voice had a low p-value (p < 0.001), and an extremely strong and positive association with ambiguous readings (0.695). This means that their joint interaction (that is, *tensepraet:voicepass*) is significant in contrast to their interactions with the ipm frequency (that is, *tensepraet:ipmVerb* and *voiceppass:ipmVerb*), which do not predict

[55] *Please Note*: The initial names of the categorical predictors used in the GLM models are *tense1*, *voice1*, *transitivity1*, and *aspect1*, where postfix 1 represents level 1 for each predictor. To simplify the visualization, 1 was replaced by the variable names these levels represent; for example, *tense1* ⇒ *tensepraet* (past tense); see Table 4.15 for complete information about the levels of the predictors.

ambiguity.

Table 4.16: Estimated regression parameters, standard errors, and p-values for the binary *GLM* Models 1, 2, and 3.

Parameter	**Estimate**	**Std. Error**	**P-value**
Intercept	-5.394	0.237	**< 2e-16** ***
ratioPtcp	1.141	0.103	**< 2e-16** ***
ipmVerb	0.796	0.07	**< 2e-16** ***
ratioPtcp:ipmVerb	0.324	0.068	**2.17e-06** ***

(a) Model 1 with ratio and ipm frequency as main effects.

Parameter	**Estimate**	**Std. Error**	**P-value**
Intercept	-2.431	0.184	**2e-16** ***
tensepraet	-0.525	0.183	**0.004** **
voicepass	0.429	0.183	**0.019** *
tensepraet:ipmVerb	-0.019	0.062	0.762
voiceppass:ipmVerb	0.022	0.064	0.737
tensepraet:voicepass	0.695	0.184	**0.0002** ***

(b) Model 2 with tense, voice, and ipm frequency as main effects.

Parameter	**Estimate**	**Std. Error**	**P-value**
Intercept	-1.967	0.076	**2e-16** ***
transitivitytran	-0.086	0.075	0.254
aspectpf	-0.008	0.075	0.919
transitivitytran:ipmVerb	0.029	0.052	0.577
aspectpf:ipmVerb	0.111	0.048	**0.021** *
transitivitytran:aspectpf	0.26	0.075	**0.001** ***

(c) Model 3 with transitivity, aspect, and ipm frequency as main effects.

Note. *Estimate*: a coefficient that indicates whether there is a positive or negative correlation between each independent variable and the dependent variable. *Std. Error* (Standard Error): a measurement of spread of the data; the larger the sample size, the close it is to the population mean, the smaller (and better) is Std. Error, *ipmVerb*: ipm frequency of verbal lemmas, *ratioPtcp*: ipm-based ratio of participial to verbal lemmas, *tensepraet*: past tense, *voicepass*: passive voice, *transitivitytran*: transitive use, *aspectpf*: perfective aspect

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (p > 0.05) is non-significant, (p < 0.05) is significant.

In Model 3, the estimated coefficients were the lowest and were negative for transitive use and the perfective aspect (-0.086 and -0.008). These coefficients increased and became positive when predictors were used in interactions: transitive use with the ipm frequency and the perfective aspect (0.029, 0.26), and the perfective aspect with the ipm frequency of 0.111. The effect of the perfective aspect was strong when it depended on the ipm frequency (p = 0.021), while the effect of the transitive use was strong when it depended on the perfective aspect (p < 0.001). Separately, transitive use and the perfective aspect did not predict ambiguity ($p > 0.1$). The positive estimated coefficients for *aspectpf:ipmVerb* and *transitivitytran:aspectpf* (0.111 and 0.26, respectively) imply that the perfective aspect interacting with the ipm frequency of base verbs and transitive perfective forms favored ambiguous readings.

Models 1, 2, and 3 were based on different datasets, which is why it was impossible to compare their estimates and to determine which model best suited the data. To investigate which of the models fitted the data best, I obtained pseudo R^2 values for each of the models and cross-checked them. A model with a higher R^2 score was expected to explain the variation in the data well. Table 4.17 shows that Model 1 had the highest R^2 score of 0.253, Model 2 had a lower R^2 of 0.025, and Model 3 had the lowest score of 0.014. Thus, Model 1 had the best fit with the data and Model 3 the poorest.

Table 4.17: *KL* R^2 scores for Models 1, 2 and 3.

	KL R^2
Model 1	0.253
Model 2	0.025
Model 3	0.014

Note. *KL* R^2: the Kullback-Leibler-divergence-based (pseudo) R^2 for generalized linear models. It measures the divergence of probability distribution of one model from another and allows the comparison of models among different datasets. The higher the pseudo R^2, the better the model fits the data.

Despite the fact that each model had a low R^2, their statistically significant p-values continue to identify that the relationships and coefficients had the same interpretation. Thus, the significance of the predictors in the models should not be discounted.

3.3.5 Test of the fit

In order to test whether the models' predictions were better than random, I resorted to a Likelihood ratio (LRT) test. An LRT test uses the maximum likelihood (for testing hypotheses concerning the variance components) to examine whether a reduced model provides the same fit as a full model. Maximum likelihood is a method that was introduced by Fisher (1922). In this method, the roles of the observed value and the distribution parameters are reversed by following the principle whereby, given a vector of observation, the LRT test compares the full model to a restricted model in which the explanatory variables of interest are omitted (without the effects in question; for example, omitting the variables of aspect, transitivity, and ipm frequency in Model 3). The p-values of the tests are calculated by comparing a null and a full model. The coefficients of the LRT test show whether there is a significant improvement in the fit of the model in comparison to the null model when predictors are added to it. The null model consists of the dependent variable *ambiguity* and only the intercept (that is, 1). The full models contain the predictors and their interactions. Using the *anova()* function with the option LRT, I computed the coefficients for the null and full models. Table 4.18 illustrates the null models (NULL, without predictors) and the full models (FULL, with the predictors and their interactions).

Table 4.18: *LRT anova* test for *GLM* Model 1, 2 and 3.

	Resid. Dev	**Df**	**Deviance**	**P-value**
NULL Model 1	2422.8			
FULL Model 1	1810.3	3	612.56	**< 2.2e-16** ***

(a) Model 1: ipm frequency, ratio.

	Resid. Dev	**Df**	**Deviance**	**P-value**
NULL Model 2	2712.5			
FULL Model 2	2645.7	6	66.822	**1.831e-12** ***

(b) Model 2: tense, voice, and ipm frequency.

	Resid. Dev	**Df**	**Deviance**	**P-value**
NULL Model 3	2165.6			
FULL Model 3	2134.3	6	31.386	**2.139e-05** ***

(c) Model 3: transitivity, aspect, and ipm frequency.

Note. *Sign. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (p > 0.05) is non-significant, (p < 0.05) is significant. *Pr(>lz|)*: the p-value. *Dev*: the deviance for the model with an intercept (null deviance) or all of the predictors (residual deviance) measuring of goodness of fit of a model (higher values indicate worse fit of the model). *Df*: Degree of Freedom which equals the number of additional parameters in a more complex model.

The LRT test for Model 1 indicated that the FULL model showed a significant improvement in terms of accuracy in comparison to the NULL model (p < 0.001, *Df* = 3). The residual deviance *DRes* was less than was null deviance (*DNull* = 2422.8, *DRes* = 1810.3), which means that the FULL model was a better fit than was the NULL model. The FULL Model 2 showed a significant improvement (compared to the null model) at the confidence level of p < 0.001, *Df* = 6. It was also a better fit than was the null model (*DNull* = 2712.5, *DRes* = 2645.7). Adding predictors to Model 3 significantly improved the accuracy of the model compared to the null model (p < 0.001, *Df* = 6, *DNull* = 2165.6, *DRes* = 2134.3). Overall, all the tested models proved to be significant predictors of ambiguity compared to the null models with the intercept as a predictor.

The significant improvement in the fit of the models (compared to the NULL models), the strong effect of the ipm frequency, and the ratios allowed me to confirm H_1 and reject H_{01}. The ipm frequency of the base verbs and the ratio of participles to their base verbs, and the dependence of the ratio on the ipm frequency of the base verbs strongly predict the ambiguity of participles. I could reject H_{02} for Model 2, as the main effects of the past tense and the passive voice, as well as the interaction thereof, proved to be significant. Rejecting H_{02} for Model 3 was more problematic because (a) the perfective aspect strongly predicted ambiguity when it depended on the ipm frequency, and (b) the transitive use strongly predicted ambiguity when it interacted with the perfective aspect. Otherwise, these factors were not significant.

3.4 Summary

The distribution analysis of ambiguous and unambiguous participles in Section 3.3.2 revealed the relationship between the ipm frequency, ratio, and ambiguity of participles. This relationship indicates that a lower rank of the base verbs (represented by their ipm frequency) accounts for a higher ratio of participial forms, many of which are ambiguous. The rank of verbal lemmas appeared to affect the proportion of participial lemmas to verbal lemmas; the higher the verbal lemma was ranked, the fewer participial lemmas it had. The number of ambiguous participles seemed to remain steady across high-frequency base verbs. The verbs of the highest rank with the ipm frequency of about 4.5 tended to have the largest number of ambiguous forms, although their ratio remained relatively low. The base verbs from the low-frequency rank formed more participles than finite/infinitival verbal forms, and among these participles, there were more ambiguous than there were unambiguous forms. In terms of the overall pervasiveness, ambiguous past passive, followed by past active participles, and transitive perfective participles had the highest mean in the distributions (Figures 4.11 and 4.12). Among these participles, there were also more ambiguous than unambiguous forms, although the total number of unambiguous participles was greater than that of ambiguous participles in the corpus (for example, 2942 versus 394 for Model 1, respectively).

The statistical analysis confirmed that the ipm frequency of verbs and the ratio of participles to verbs had a significant effect on ambiguity (Model 1). According to the estimated coefficients in Model 1, the base verbs with higher ipm frequencies and higher ratios were likely to form more ambiguous participles. Contrary to the coefficient for the ipm frequency, Figure 4.9 indicates that the base verbs, except for the most frequent verbs, formed more ambiguous participles as their rank (expressed by the ipm frequency) decreased. In this figure, numerous participial forms clustered around the most frequent interval 4–4.7 and might have biased the high score of the coefficient. For this reason, given the observations of the entire distribution of the ambiguous forms, the significance of the relationship between higher ipm frequency and ambiguous forms should be considered with caution. The higher ratio of participles implies that it is more likely for a participle to be ambiguous when the base verb have more participial than finite/infinitival verbal forms; this holds true both for the base verbs with low frequency and those with the highest frequency (that is, in the interval 4–4.7).

The analysis also shows that the past tense and the passive voice explain ambiguity, both alone and when the past tense depends on passive voice. Unlike the passive voice, which is strongly associated with ambiguous readings, the past tense alone favors unambiguous readings (Model 2). The relationship between past passive participles and ambiguity is the strongest, although it does not depend on the frequency of the base verbs. Use of transitive forms and the perfective aspect cannot predict ambiguity alone; however, (a) interaction between the perfective aspect and ipm frequency and (b) the transitive use conditioned by the perfective aspect showed a stronger effect on ambiguity (Model 3). The LRT test confirmed the significance of these predictions (compared to the intercept) across all the models. The best fit of Model 1 suggests that the ipm frequency and ratio are more robust than the remaining models in predicting ambiguity.

The results of the statistical analysis concur with the morphosemantic approach. Past passive participles tend to favor adjectivization because of the resultative meaning of the past tense and the reduced argument structure of passive forms. In addition, the passive voice renders the past tense more resultative and less temporal, the past tense disfavors adjectivization because of its more defined temporal meaning (Table 3.3). The clear evidence from Model 3 (see Table 4.18c, Figure 4.12) demonstrates that, in this study, perfective transitive forms explain ambiguous uses of participles, across the frequencies of their base verbs; this is not surprising because (a) transitive perfective forms are common and form two types of participles, see Table 4.8, (b) transitive perfective forms have the highest mean ratios for ambiguous participles (TR PFV in Figure 4.12), and (c) the perfective aspect conveys a resultative meaning of action and favors adjectivized uses (Kalakuckaja, 1971; Kolochkova, 2011).

The significance of the interaction between transitive use and the perfective aspect can be related to Hopper and Thompson's Transitivity Hypothesis. In this hypothesis, aspect is systemically correlated with the degree of transitivity. The perfective aspect is associated with a more transitive use, and the imperfective aspect is associated with a less transitive use (as in Finnish). Russian linguists specializing in aspect support this claim, but for them, aspect is defined in terms of telicity (telic or atelic) rather than perfectivity and the presence or absence of an object in transitive or intransitive verbs (Rassudova 1982; Forsyth 1970, as cited in Dunlap 1981).

4 Conclusion

The exploratory analysis of participial and verbal word forms across corpus data highlighted the properties of semantic and morphosyntactic factors that can account for adjectivization. The most relevant factors of adjectivization proved to be the pervasiveness of participles (expressed in their ratio), the frequency rank of the base verbs, the past tense, and the passive voice. The relevance of the adverb of measure and degree *očen′*, as a factor of adjectivization, is provisional because a syntactic context should be considered first.

The results of the first corpus study indicate that *očen′*, as a syntactic factor of adjectivization, appear to be marginal in the *očen′* + PRESP construction in the Araneum corpus and was observed more frequently with finite verbs than with participles. The significant association between ratio and type of construction (finite verbal or participial) implies that participles, in comparison to finite verbal forms, are not typically used with *očen′*. The significant association of the semantic classes with the ratio of the *očen′* participial constructions[56] supports the assumption in which a gradable semantic component of a verbal (finite or participial) or adjectival form allows the combination with *očen′* (Sičinava, 2018; Lundquist et al., 2013). Thus, the use of a participle with *očen′* signals that its semantics, inherited from its base verb, permits this use. Furthermore, the class of psychological and mental domains accounts for the highest mean ratio of the *očen′* + PRESP construction. The qualitative analysis of the eight representative examples showed that the participles of these semantic classes, which were used with *očen′* more often than their corresponding finite verbs were used, were adjectivized. While this observation does not relate the semantic classes to adjectivization of the participles in this study, the role of semantics and the associated ratio may have some effect on predisposing a participle towards adjectival uses with *očen′*. Syntactic context plays a primary role in distinguishing between adjectivized and unambiguous uses of participles with *očen′*: If a participle with the special semantic class combines with *očen′* and its syntactic context is devoid of verb complements and adjuncts, then this word form may be adjectivized. If, on the contrary, the context contains complements and adjuncts, then the word form is likely to be an unambiguous participle. The use of a participle with *očen′* may serve nevertheless as a criterion of adjectivization only if the syntactic context exhibits other signs of adjectivization.

The second corpus study demonstrates a significant relationship among several morphosyntactic factors of adjectivization and lexical frequency of participles in the disambiguated subcorpus of the RNC, and ambiguity of participles. The observations of three ranges of the frequency distribution showed that the high- and low-frequency verbal lemmas were the most representative in the pervasiveness of participles among the other verbal forms. High-frequency verbs formed more finite/infinitival verbal forms than participles did, but they formed participles continuously; underscoring this trend was the lower percentage of high-frequency verbs with no participles relative to that of mid- and low-frequency verbs. Approximately one-third (35.5%) of all low-frequency verbs tended to form more participles than they formed finite verbs and infinitives, while only 9.4% of these verbs formed more finite verbs and infinitives than they formed participles. The analysis also showed a relationship between the pervasiveness of morphological types of participles and their ambiguity in terms of a shift in lexical meaning. The examples from the frequency ranges indicate that high-frequency present active participles and low-frequency past passive participles tend to be adjectivized (with an optional extension of meaning to figurative uses); for example, the adjectivized participles derived from the high-frequency base verbs tend to gain additional meaning; that is, *govorjashij slon* ‘talking elephant’ versus *govorjashaja familija* ‘self-explanatory surname’.

The statistical analysis complemented and confirmed some of the observations made in the analysis of the frequency distribution. The rank of the base verbs and the pervasiveness of participles are powerful predictors of ambiguity, in comparison to the morphosyntactic factors of

[56] This also applies to the finite verbal constructions.

adjectivization. The higher ipm frequency of the base verbs and the higher ratio of participles strongly predict ambiguous participles, although the analysis of frequency ranges demonstrates that low-frequency verbal lemmas and the verbal lemmas clustering on the interval 4–4.7 had the highest ratio of ambiguous participles. In terms of ratio, higher-ranked verbal lemmas formed fewer corresponding participial lemmas than lower-ranked verbal lemmas formed; this observation suggests that the frequency of the base verbs affects the pervasiveness of participles, and this pervasiveness also determines how many participles can be adjectivized.

Morphological types of participles, such as past passive and transitive perfective participles, also show a strong relationship with ambiguous forms, which supports the claim of the morphosemantic approach in which internal grammatical properties of participles relate to the development of adjectival properties within a participial lexeme. As distinct categories, the past tense and the passive voice are strong predictors of ambiguity: The past tense explains unambiguous use, while the passive voice elucidates the use of ambiguous forms. Past passive participles also show a strong relationship with ambiguity; Kustova (2012) refers to them as one of the largest groups of adjectivized participles (see also Kalakuckaja 1971). The syntactic properties of tense, voice, aspect, and transitivity that affect adjectivization (that is, reduced argument structure, verb government, temporal meaning, or concrete meaning of the base verb) underlie the ability of these factors to predict ambiguity. Aspect did not prove to be statistically significant for explaining ambiguous uses as a separate category, but the perfective aspect's interaction with the ipm frequency of the base verbs had a strong effect on ambiguous participles. According to the results of Model 3, the transitive use did not predict ambiguity and was more associated with unambiguous participles, but perfective transitive participles held a significant relationship with ambiguous forms, which implies a close correlation between the perfective aspect and the degree of transitivity of the verb (Dunlap, 1981). Most interestingly, the samples of the verbs that did not have corresponding participles contained many intransitive perfective forms, which demonstrate the fewest number of morphological types of participles in grammars (Table 4.8). Therefore, I argue that the group of intransitive perfective verbs has such a low chance of forming participles frequently that these participles might be adjectivized.

Chapter 5

Disambiguation

1 Introduction

In this chapter, I develop a disambiguation model that aims to resolve the ambiguity between participles and adjectives as a result of adjectivization. This ambiguity is reflected in the output of the morphological analyzer, which assigns double participial/adjectival readings to adjectivized participles. The model is expected to resolve ambiguity by analyzing ambiguous participles as being either participles or adjectives. It relies on the morphosyntactic factors of adjectivization discussed in Chapter 3, and the lexical frequency of POS lemmas for participles and adjectives that are transformed logarithmically into weights.

In the chapter, I approach adjectivization as one of the persistent ambiguity problems in the Russian corpus (for example, 24.10% in SynTagRus, as discussed by Klyšinskij and Rysakov 2015), whereby ambiguous participles need to be distinguished from adjectives using formal methods. To achieve this, I will apply the factors of adjectivization outlined in Chapter 3 and the lexical frequency of adjectival and participial lemmas as the basis for the rule-based approach implemented in this dissertation.

The motivation for making the participle-adjective distinction lies in the experimental design, which allows one to process the properties of adjectivized participles at both fine- and coarse-grained levels, with a major focus on the syntactic context, internal morphological features, and corpus frequencies represented by weights. The research questions that I aim to answer by examining the performance of the disambiguation model are as follows:

- How can frequency assist in the resolution of POS ambiguity?
- How well can the syntactic and morphological factors of adjectivization resolve the ambiguity of participles?
- Does disambiguation work better when weights and factors of adjectivization are used together or separately?

The evaluation of the disambiguation model is intrinsic; that is, it applies to a specific subtask of distinguishing participles from adjectives as part of the main task of POS ambiguity resolution and weighting. This subtask does not apply to persistent ambiguity, such as intra-paradigmatic ambiguity or POS ambiguity with unrelated morphological forms and meanings. However, it involves several steps based on linguistic material, such as the syntactic context, morphological properties, and corpus frequencies. The design of the experimental design is novel, and may be further implemented as a distinct method in CG rules for other languages or within the Russian CG for addressing other linguistic phenomena. Unlike the use of typical CGs and weights for highly inflected languages that manage ambiguities based on affixes and inflection, the disambiguation of adjectivized participles focuses on syntactic properties and weights by using internal morphological properties (including affixes) as supplements. For this reason, word order, syntactic categories, and the constituents of the immediate syntactic context are essential components in the design of the CG rules.

The model, the development and implementation of which I will present and explain in this chapter, is based on several methods. The first method is weighting; that is, implementing the corpus frequencies of participial and adjectival lemmas as weights in the morphological analyzer for Russian. The second method involves building a customized gold-standard corpus (also known as the gold standard, and which is based on written texts) that is linguistically motivated and applies specifically to one type of ambiguity. When constructing the gold standard, I ensured that adjectival and participial word forms, together with the morphological types of participles, were proportional to each other. Moreover, the most problematic cases of ambiguous word forms were annotated by a larger number of Russian respondents, in addition to the clear-cut adjectives and participles that I annotated manually. The third technique concerns differentiating among the CG rules based on the syntactic context, the morphological properties, and weights, either separately or combined with the morphological and syntactic rules. The syntactic context and morphological properties are supported by the theoretical approaches to adjectivization[1] that take all the morphosyntactic properties of adjectivized participles into consideration. The end result is a CG grammar that operates by using linguistic information in a comprehensive manner and relies on the weights. The weighting method could eventually be extended to the evaluation of larger corpora and tested on several linguistic phenomena related to POS ambiguity.

The chapter is structured as follows: Section 2 illustrates the weighting methods, and presents the transformation and implementation of weights in the Russian morphological analyzer. Section 3 discusses the development of the gold standard for disambiguation, and focuses on the annotation of the most difficult cases identified in the gold standard by conducting a survey experiment. Section 4 describes the design of the CG rules and the disambiguation submodels that cover specific types of CG rules. Section 5 follows with a description of the performance of each submodel in the disambiguation experiment and an interpretation of the results, while Section 6 concludes with the presentation of the results of the disambiguation model and weighting.

[1] See Chapter 3, Sections 4.1 and 4.2.

2 Weighting

In this section, I discuss the methodological background to weights and the development of weighting; that is, transforming frequency values into a transducer-readable format, primarily for participles, finite verbal forms, and adjectives. First, I review existing methods that implement weights for disambiguation that have been discussed in several studies. Second, I introduce the concept of weight as used in this dissertation, and the corpus data upon which it is based. Finally, I present the methodology and implementation of weights in the Russian finite-state transducer (FST).[2] Weighting is an important component of disambiguation because it allows one to assess how well weights can be used in disambiguation tasks, either alone or combined with the rules based on linguistic properties. As demonstrated previously in Chapter 4, Section 3.3.3, corpus frequency distribution has a significant effect on the ambiguity of participles and may consequently reflect general mechanisms behind POS categories (Piantadosi, 2014). From this perspective, weights, dependent on the corpus frequency distributions of their respective lemmas, can function as a formal criterion for distinguishing between adjectives and participles, similar to the morphosyntactic factors of adjectivization.

2.1 Background to weighting

The section focuses on studies that offer the methods that I adapt for weighting the Russian FST. The main goals are to define weights and implement them in the Russian FST.

2.1.1 Implementation of weights

The implementation of weights in this dissertation is largely based on the method applied by Lindén et al. (2009a) and Linden and Pirinen (2009). The method involves the unigram statistics (frequency-based statistics for word forms) used for selecting the most likely morpheme segmentation and the most frequent reading of each compound word form in Finnish. Lindén et al. (2009a) weighted different parts of the lexicon with frequency data from a corpus using weighted finite-state transducer calculus, and then ran the weighted morphological analyzer on Finnish text. As a result, the weighted lexicon and unigram statistics proved to be beneficial[3] for disambiguating Finnish compound word forms and their segmented morphemes, because Finish is morphologically complex and requires little or no context for disambiguation.

The core parts of the Lindén et al.'s (2009a) approach are the estimation of probabilities (Equation (5.1)) and the weighting of the lexicon (Equation (5.3)). Equation (5.1) shows that the probability p of a token a or $p(a)$ occurring in the corpus is computed by dividing the count of a

[2] *Please note*: In this chapter, the Russian FST is the morphological analyzer discussed in Chapter 2, Section 3.4.
[3] The WFST achieved 96–98% precision for Finnish compounds in the vocabulary.

token $c\,(a)$ by the corpus size cs.

$$p\,(a) = c\,(a)\,/\,cs \tag{5.1}$$

The count of a token $c\,(a)$ is defined as its corpus frequency plus 1; that is, Laplace smoothing, which takes missing word forms or lemmas in the corpus into account; see Equation (5.2).

$$c\,(a) = 1 + \text{frequency}(a) \tag{5.2}$$

Equation (5.3) indicates that the weights of tokens $w(a)$ are implemented in the tropical semiring using the negative log-probabilities of the token $-log(p(a))$. As a result, the lowest corpus weight is associated with the most likely reading, and the highest corpus weight with the most unlikely reading of the original lexical transducer.

$$w(a) = -\log\,(p\,(a)) \tag{5.3}$$

The Lindén et al.'s (2009a) approach reflects earlier works based on the estimation of probabilities for disambiguating compounds in inflected languages such as German (Schiller, 2005; Marek, 2006) and Swedish (Karlsson, 1992; Sjöbergh and Kann, 2004). In a more recent work in 2020, Keleg et al. (2020) implemented WFSTs that were based on unigram counts (as in Linden and Pirinen 2009) and on the CG counts and *word2vec* vector-based counts (for the semantics of the words) in order to resolve the POS ambiguity (for example, *wound* as a verb versus as a noun) in English, Serbo-Croatian and Kazakh. The unigram count-based method using Laplace smoothing had the most accurate performance, with accuracy above 80% compared to the constraint-based and *word2vec* (neural network) methods.

2.1.2 Transformation of weights

Weights encode log-transformed corpus frequencies of word forms, and convey the effect of the word-frequency distributions (Piantadosi, 2014, 2012; Dehaene and Mehler, 1992; Calude and Pagel, 2011). In order to improve the interpretation of the word-frequency effect, a simplification measure can be applied to the weight values. One of the standardized measures that could be applied to word frequencies was suggested by van Heuven et al. (2014). The measure reduces differences in word-frequency counts among word forms across large corpora, and makes word-frequency values more comparable. The simplification aims to manage absolute numbers representing counts of words across large corpora, which are difficult to interpret. This measure overcomes the drawbacks of the standardized measure of ipm (instance per million words) frequency, such as a frequency effect below 1 ipm[4] and an insignificant difference between the

[4] That is, corpora that is larger than 1 million words (for example, 100 million or 100 billion words) leads to the situation in which numerous types have frequencies below 1 ipm.

means of differences of words below and above 1 ipm.[5]

Equation (5.4) illustrates the Zipf measure introduced by van Heuven et al. (2014)], which converts the values of ipm frequencies into more understandable values on a scale from 1 to 7.

$$\text{Zipf} = \log_{10}\left(\frac{w+1}{c+wt}\right) + 3.0 \tag{5.4}$$

The values on the Zipf scale are calculated by log-transforming the frequency values per million words[6] $\frac{w+1}{c+\text{wt}}$ and scaling them by adding 3.0. w is the count of a word in a corpus, $w+1$ is the Laplace smoothing, c is the size of a corpus in millions, and wt is the number of word types in millions. Thus, $\log_{10}(\dots) + 3.0$ logarithmically transforms frequency values per million words and adds them on a scale from 1 to 7.

Table 5.1 illustrates the Zipf scale (described above) ranging from 1 to 7. Words with Zipf values of 3 or lower are low-frequency words, while words with Zipf values of 4 and higher are high-frequency words. The value of 6 represents high-frequency content words, while that of 7 represents a few function words, pronouns, and verb forms such as *have*. The interval between 3 and 4 is the extreme from low to high frequency (I will refer to it henceforth as *medium frequency*). The examples in the table are based on the SUBTLEX-UK[7] word frequencies.

Zipf scale	**IPM frequency**	**Examples**
1	0.01	*antifungal*, *bioengineering*, *farsighted*, *harelip*, *proofread*
2	0.1	*airstream*, *doorkeeper*, *neckwear*, *outsized*, *sunshade*
3	1	*beanstalk*, *cornerstone*, *dumpling*, *insatiable*, *perpetrator*
4	10	*dirt*, *fantasy*, *muffin*, *offensive*, *transition*, *widespread*
5	100	*basically*, *bedroom*, *drive*, *issues*, *period*, *spot*, *worse*
6	1000	*day*, *great*, *other*, *should*, *something*, *work*, *years*
7	10,000	*and*, *for*, *have*, *I*, *on*, *the*, *this*, *that*, *you*

Table 5.1: The Zipf scale of word frequency (van Heuven et al., 2014: 1180).

2.2 Weighting the Russian finite-state transducer

In this section, I introduce my own approach to weighting based on the set of methods adopted from Lindén et al. (2009a), Linden and Pirinen (2009) and van Heuven et al. (2014). The methods reflect the notions of probability of a lemma in a corpus and frequency rank of a lemma in a corpus. The probability of a lemma corresponds to the logarithmic transformation of corpus frequencies of verbal and adjectival lemmas, representing the probability of a given lemma in a corpus (Lindén et al., 2009a; Linden and Pirinen, 2009). The frequency rank of a lemma implies

[5] For example, the mean of 1 ipm and 0.7 ipm is 0.5 ipm, and the mean of 5 ipm and 35 ipm is 3 ipm (van Heuven et al., 2014).

[6] For transforming frequency values per billion words, just $\log_{10}(\dots)$ can be used.

[7] SUBTLEX-UK is a word-frequency database based on subtitles of British television programs.

that its frequency is inversely proportional to its frequency rank. It also refers to the scaling of log-transformed frequency values into weights according to a frequency ranking of lemmas (low frequency, medium frequency, and high frequency; van Heuven et al. 2014).

In this approach, *weight* signifies the corpus frequency of a lemma, logarithmically transformed and conveying the probability of that lemma's use in the lexicon of a language. It conveys the frequency-wise importance of a lemma in a corpus in relation to the effect of the POS of a lemma on its corpus frequency. To give an example, Dehaene and Mehler (1992) and Piantadosi (2012) show that the frequency of number words ("one", "two", "three", and so on) is predictable from the cardinality to which the words refer and is also cross-linguistically predictable. The assumption that the linguistic information is encoded in corpus frequencies of lemmas and can be significant for identifying their respective POSs (Piantadosi, 2014, 2012; Dehaene and Mehler, 1992) is adopted in this dissertation as the basis for weighting.

I adopt Equation (5.3)[8] as a base for weighting, establishing slightly different parameters for the mode of weight use in the Russian FST. First, the values of the weights convey positive log probabilities; that is, the highest weight will denote the highest probability (frequency), and the lowest weight will signify the lowest probability. Second, the weights encode log probabilities of lemmas instead of tokens (or word forms). Lemma types represent lexemes for a given token, such as *govorila* '[she] was saying-V.PST.F ', *govorit* '[he] says-V.PRS.M', *govorjaščij* 'talking-PRESP.M'. These are tokens, whereas *govorit′* 'say, talk' is a lemma type. The difference between the frequency of a token and that of a lemma may differ across the corpus; this is why adding one as a missing token in Laplace smoothing will not necessarily work the same way for lemmas, as one missing token does not imply one missing lemma.

2.2.1 Data description

I obtained lemmas and their frequencies from Sharoff's frequency list of lemmas.[9] The list was compiled on the basis of the Russian Internet Corpus (I-RU; Sharoff 2006; Sharoff et al. 2013), annotated using the Trigrams'n'Tags (TnT) tagger (Brants, 2000; Sharoff and Nivre, 2011) and the Center for SprogTeknologi[10] (CST) lemmatizer (Sharoff and Nivre, 2011). The TnT tagger is a probabilistic tagger based on Hidden Markov Model (HMM), approximating the probabilities of unseen tag sequences, and guessing possible tags of unseen words. It is trained on the manually disambiguated subcorpus of the RNC and uses the unified tagset MULTEXT-East (MTE;[11] Sharoff et al. 2008).

I-RU is based on more than 30,000 web-pages: total number of word forms is 147,803,971 (words with Cyrillic characters); total number of lemmas, 1,078,346 (orthographic Cyrillic-only lemmas in the lexicon of this corpus; Sharoff et al. 2013), see Table 5.2.[12]

[8] Equation 5.3 is used for implementing corpus probabilities as weights in the tropical semiring.

[9] Available at: http://corpus.leeds.ac.uk/frqc/internet-ru-lpos.num.xz

[10] Center for Language and Technology.

[11] Available at: https://www.sketchengine.eu/russian-tagset/

[12] *Please Note*: There are some inconsistencies in the estimates of the I-RU corpus data (number of word forms

Components	#
web-pages	30,000
words	147,803,971
lemmas	1,078,346

Table 5.2: Estimates for the I-RU lexicon for components of the corpus.

Table 5.3 indicates that each lemma has a major category (*V* for a verb, *A* for an adjective, *N* for a noun) and their features at fixed position (such as type, verbal form, tense, gender, number, person, and so on). For example, *Vmis-sfa-p* for the lemma *подтвердить* 'confirm' stands for verb, main, indicative, past, -, singular, feminine, active, -, progressive. This analysis applies to a particular type of the verb in question represented by *Vmis-sfa-p* set of tags (among other verbal types which hold the same set of tags).

Rank	Frequency	Lemma	Tag
16949	884	*настоящий*	Afpmsdf
16955	884	*фабрика*	Ncfsnn
16958	884	*подтвердить*	Vmis-sfa-p

Table 5.3: The structure of the frequency list including the frequency rank, raw frequencies, the lemmas *настоящий* 'real', *фабрика* 'factory', *расходиться* 'part', *подтвердить* 'confirm', and their tags ordered by the frequency rank of the lemmas.

2.2.2 Transformation of weights

Turning corpus frequencies into weights requires a logarithmic transformation as shown in Lindén et al. (2009a) and Linden and Pirinen (2009). In order to refine the values and make them comparable to their corpus rank, I adopted the technique of combining logarithmic transformation and scaling. Scaling is necessary for establishing a threshold for weights based on their frequency rank in a corpus according to low-frequency, medium-frequency and high-frequency lemma types. The threshold can be represented by a Likert-type scale ranging from rank classes of higher and lower corpus frequencies.

To address lemma types, I readjusted Equation (5.4), including the notation, by replacing types with lemmas. Equation (5.5) consists of the normalization of ipm corpus frequencies, with smoothing and scaling as the main parameters.

$$\begin{aligned} \text{Zipf} &= \log_{10}\left(\frac{\text{rawFreqSmooth}}{\text{corpusSizeIpm} + \text{lemmasNbIpm}}\right) + 3.0, \\ \text{rawFreqSmooth} &= \text{rawFreq} + 1 \end{aligned} \tag{5.5}$$

and lemmas) presented in Sharoff (2006) and Sharoff et al. (2013). I adopted the estimates in Sharoff et al. (2013).

Normalization is represented by $\left(\frac{\text{rawFreqSmooth}}{\text{corpusSizeIpm}+\text{typesNbIpm}}\right)$, Laplace smoothing is rawFreq $+$ 1. rawFreq is the corpus frequency of a lemma, corpusSizeIpm is a total number of tokens in a corpus per million, and lemmasNbIpm is the total number of lemmas per million.[13] $\log_{10}\left(\frac{\text{rawFreqSmooth}}{\text{corpusSizePm}+\text{typesNbPm}}\right)$ represents the Zipf smoothing; that is, reducing the difference between per million frequency values; $\log_{10}$ is a function smoothing the difference between the raw frequency values of lemmas in the corpus, and 3 is the scaling of the value output from $\log_{10}\left(\frac{\text{rawFreqSmooth}}{\text{corpusSizePm}+\text{typesNbPm}}\right)$ on the interval from 1 to 7, which prevents the lowest frequency values from becoming negative.

Table 5.4 illustrates how the raw frequency values of the lemmas changed with the transformational method on the basis of Equation (5.5): Laplace smoothing, Zipf smoothing, and scaling.[14] The cell *rank* represents the weight values scaled according to high-, medium-, and low-frequency ranks.

Lemma	**Raw frequency**	**Laplace smoothing**	**Zipf smoothing**	**Scaling (weight)**	**Rank**
называть 'name'	17463	17464	2.07	5.07	high
открыть 'open'	6528	6529	1.64	4.64	medium
отказать 'reject'	599	600	0.61	3.61	medium
забанить 'ban'	78	79	-0.28	2.72	low
привинтить 'screw'	10	11	-1.13	1.87	low

Table 5.4: Types of value transformation for the lemmas that correspond to participial word forms in the verbal lexicon.

Table 5.4 indicates that Zipf smoothing decreased the differences in the lemma frequencies to a minimal degree. For example, the difference between the raw frequencies for 'name' and 'open' is 10935, whereas the difference between the frequencies after Zipf smoothing is 0.43.[15] The frequency values are minimally different when they are scaled compared to raw corpus frequencies. As shown in Table 5.4, the scaling placed smoothed values on the Zipf scale from 1 to 7 in which the weight value of 5.07 represents high-frequency lemmas, and 4.64 and 3.61 medium-frequency lemmas, and the remaining weights, 2.72 and 1.87, low-frequency lemmas. I did not detect any lemmas for participial word forms with frequencies corresponding to the rank of 6 in Sharoff et al.'s (2013) frequency list.

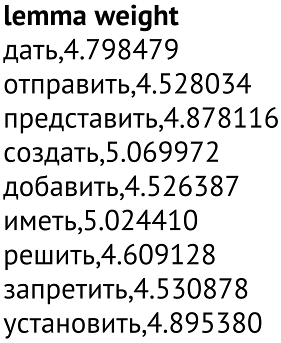
[13] The ipm method consists of normalizing the number of occurrences of an item per million; that is, dividing the total number of occurrences of an item by the corpus size.

[14] *Please Note*: Apart from raw frequency values, all the other values were rounded to two decimal places in order to simplify their visualization.

[15] The difference values were computed in the following way: $17463 - 6528 = 10935$, $2.07 - 1.64 = 0.43$.

2.2.3 Implementation of weights

First, I retrieved the lemmas and their corresponding raw frequencies with participial (*Vmp...*), finite verbal, (*Vmi...*),[16] and adjectival tags (*A...*) from Sharoff et al.'s (2013) frequency list. Second, I transformed the lemma frequencies into weights using Equation (5.5), discussed in Section 2.2.2. This equation allows one to transform raw corpus frequency values into the Likert-scale values from 1 to 7 (called "The Zipf scale of word frequency"), in which Zipf values from 1 to 3 represent low-frequency words, values between 3 and 4 refer to medium-frequency words, and values of 4 to 7 represent high-frequency words. Equation (5.5) was implemented in Python script, which transforms raw frequency values into weights and stores them in separate lists for participles and adjectives. Figure 5.1 illustrates a sample of the list containing participial lemmas retrieved from Sharoff et al.'s (2013) frequency list and their corresponding weight values.[17] For example, the verbal lemma *дать* 'give' has a weight of 4.798479, the verbal lemma *отправить* 'send' has a weight of 4.528034. I then used the lists of participial and adjectival lemmas to assign weights to verbal and adjectival lemmas in the lexicons of the analyzer. Finally, all of the lemmas from these lists were compared to the lemmas in the verbal and adjectival lexicons.[18]



```
lemma weight
дать,4.798479
отправить,4.528034
представить,4.878116
создать,5.069972
добавить,4.526387
иметь,5.024410
решить,4.609128
запретить,4.530878
установить,4.895380
```

Figure 5.1: A sample of the list consisting of participial lemmas (1st column) and their corresponding weights (2nd column) obtained after the Zipf transformation of the corpus frequencies.

The weighted lexicons can provide weights for the output of the morphological analysis. Table 5.5 below shows the morphological analysis and the weights for the ambiguous word form *курящий* /kurjašij/'smoking'.

[16] I did not use the lemmas for finite verbal word forms for disambiguation in this dissertation; however, their weights are available in the file https://github.com/giellalt/lang-rus/blob/develop/src/fst/stems/wverbs.lexc

[17] *Please Note*: The weight values in this sample were rounded to six decimal places.

[18] See Chapter 2, Figure 2.8

Word form	Morphological analysis	POS	Weight
курящий	*курящий*+N+Msc+Anim+Sg+Nom	N	0.000000
	курить+V+Impf+IV+PrsAct+Msc+AnIn+Sg+Nom	PTCP	3.144531
	курящий+A+Msc+AnIn+Sg+Nom	A	3.279297

Table 5.5: The structure of morphological readings with weights assigned by the morphological analyzer. The structure consists of a word form, a morphological analysis of each POS, and the weight associated with it.

The nominal reading курящий+*N*+*Msc*+*Anim*+*Sg*+*Nom* has a weight of 0.000000, the verbal readings курить+*V*+. . . have a weight of 3.144531, and the adjectival readings курящий+*A*+. . . have a weight of 3.279297. The digits after the decimal place represent the difference in weight values, while the first digit before the decimal place indicates that both verbal and adjectival lemmas belong to the low-frequency rank of 3.

The percentage of weighted and unweighted lemmas in the verbal and adjectival lexicons is presented in Figure 5.2.[19]

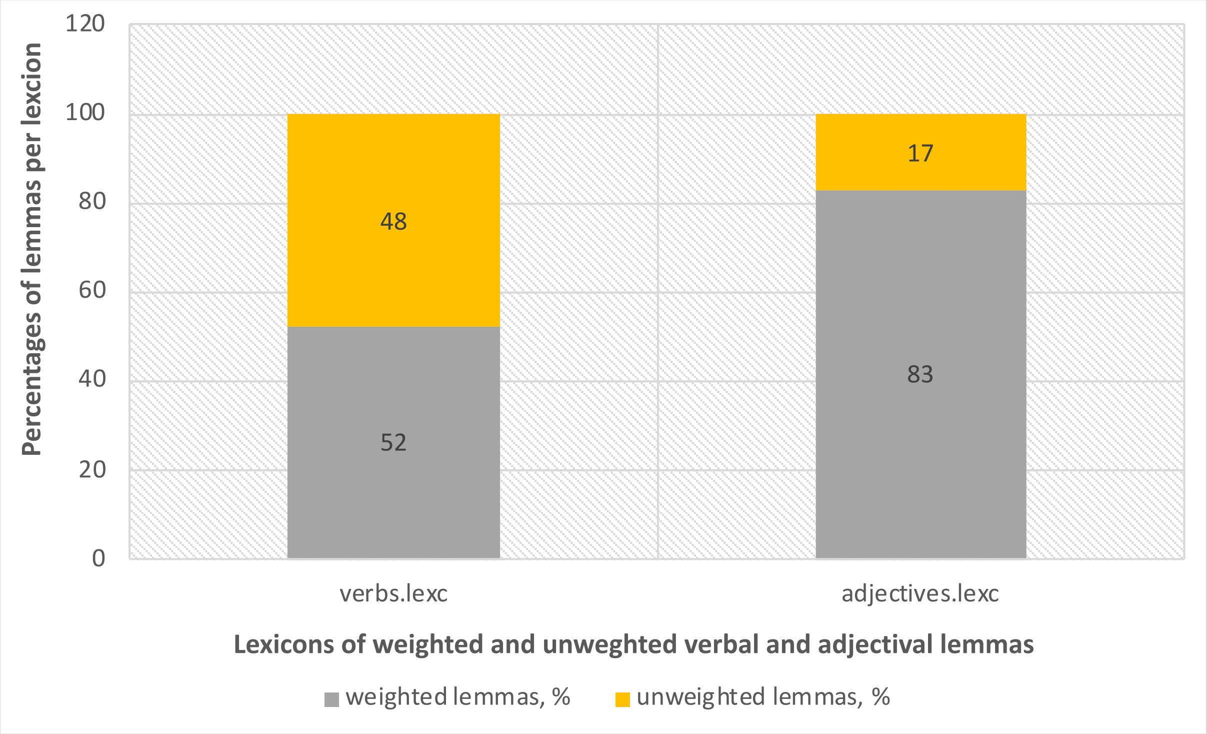


Figure 5.2: Distribution of weighted and unweighted lemmas across the lexicons *verbs.lexc* and *adjectives.lexc*.

In the verbal lexicon, the percentage of weighted lemmas is slightly higher than is that of unweighted ones (52% versus 48%). In the adjectival lexicon, the percentage of weighted lemmas accounts for 83% of all entries, compared to 17% for unweighted lemmas. As the weighted

[19] For each lexicon, the percentage was calculated by dividing the number of weighted/unweighted lemmas by the total number of lemmas in the given lexicon multiplied by 100.

lemmas in the verbal lexicon only represent participial word forms, their percentage is lower than is the percentage of the weighted adjectival lemmas.

2.2.4 Summary

In the approach to weighting the Russian FST, weights represent logarithmically transformed and scaled corpus frequencies of participial and adjectival lemmas taken from Sharoff et al.'s (2013) frequency list. Frequencies of the lemmas are based on the summed frequencies of the types annotated by the probabilistic tagger trained on the disambiguated subcorpus of the RNC, so these frequencies should reflect the distribution of tags in this subcorpus.

The lemmas represent only the grammatical category of word form as a participle or an adjective. Their frequencies are atomic and do not account for a morphological type of participle, such as a past/present active/passive participle. Thus, all of the verbal lemmas are annotated with weights, based on the frequency of participles, and all of the adjectival lemmas are annotated with weights based on the frequencies of adjectives. I did not implement weights for types of participles, as the structures of the lexicons do not allow individually weighting each verbal lemma and its participial inflection. This is because all of the continuation lexicons represent inflection types for different groups of participial lemmas, making it impossible to assign weights for inflections in individual lemmas.

The transformation is performed by means of the standardized measure adopted from van Heuven et al. (2014); Lindén et al. (2009a); Linden and Pirinen (2009), which allows one to reduce the difference between different weight values and put them on the frequency scale, from the least frequent (1–3) to the most frequent lemmas (4–6). I assigned weights to the lemmas in the verbal and adjectival lexicons to enable disambiguation with weighted rules.

3 Development of the gold standard

This section focuses on the development of the gold-standard for the following disambiguation experiment. The importance of this task lies in the overall objective of this chapter: creating a CG that facilitates a distinction between ambiguous participles and adjectives using information from the syntactic context, morphosyntactic properties, and weights. For this reason, the design of the gold standard is largely based on the balanced distribution of the morphosyntactic properties of the participial word forms and the properties of their immediate syntactic context. As I explore only one type (that is, the POS ambiguity of participles), and I use the framework of CG, I do not need to create a large reference set for testing the CG rules.

3.0.1 Data description

The sentences used in the gold standard were taken from the SynTagRus corpus.[20] SynTagRus is a dependency treebank (deeply annotated corpus of Russian) with morphological/syntactic annotation, humanly corrected, based on morphological and syntactic annotation of Universal Dependencies (UD[21]). SynTagRus is fully disambiguated (morphologically and syntactically); every word form is assigned one POS tag and a unique set of morphological features. SynTagRus is a subcorpus of the RNC and currently contains over 1,000,000 tokens (over 66,000 sentences; see Table 5.6) belonging to texts from a variety of genres such as contemporary fiction, popular science, newspaper and journal articles dated between 1960 and 2016, texts of online news, and so on.

Components	#
sentences	> 60,000
tokens	> 1,000,000

Table 5.6: Estimates for the components of the SynTagRus dependency treebank.

I resorted to SynTagRus primarily because it was annotated semi-automatically (processed) by the ETAP parser[22] and then manually corrected by linguistic experts, so there was an expectation of a certain annotation quality due to the combination of automatic and manual annotation.

3.0.2 Selection of sentences for the gold standard

To compile the gold standard for disambiguation, I extracted the sentences (running text) containing instances of participles and adjectives[23] from the SynTagRus corpus. I then shuffled these sentences randomly, selecting approximately 295 sentences. Among these sentences, instances of participles were used in heterogeneous syntactic contexts and were derived from different verbal lemmas. The sentences were taken from the SynTagRus corpus as is and were not edited, artificially constructed, or shortened.

First, I manually selected sentences that included participles with the double reading ADJ/PTCP output by the analyzer. Then I disambiguated only the word forms in which (a) syntactic context clearly pointed at their adjectival or participial properties, or (b) meaning was clearly verbal or adjectival. I resorted to the factors of adjectivization (such as complements, adjuncts, adverbs of measure/degree, preposed/postposed position) to check context. I also used my native speaker's

[20] Available at: https://github.com/UniversalDependencies/UD_Russian-SynTagRus

[21] Available at: http://universaldependencies.org/introduction.html

[22] ETAP or ETAP-3 is a morphological analyzer using a morphological dictionary to produce morphological annotation of word forms from the corpus; the annotation includes lemmas, POS tags, and a set of morphological features for each POS.

[23] I searched for instances of participles and adjectives using the tags *VerbForm=Part* (for participles) and *ADJ* (for adjectives).

intuition as a support criterion for annotating the word forms, especially when the context did not exhaustively indicate if a word form was a participle.

The remaining, seemingly ambiguous word forms (with or without double ADJ/PTCP readings) that I could not disambiguate myself were suggested to a larger group of native Russian speakers. More specifically, the sentences with these word forms were added as examples to the survey (see Section 3.0.3), in which 43 Russian respondents selected a verbal or adjectival interpretation for an ambiguous word form in each example.

Table 5.7 gives an overview of the counts for word forms and sentences in two components of the gold standard. *Primary annotation* concerns the clear-cut cases of adjectives and participles that I annotated. *Survey annotation* covered the problematic cases that constitute about 15% of all the gold standard (in addition to 5% of the control sentences) and were annotated by a large group of native speakers.

Items	**Primary annotation**	**Survey annotation (incl. /excl. controls)**
word forms	202	51/48[24]
sentences	221	50

Table 5.7: Counts of word forms and sentences in two subparts of the gold standard, primary annotation and survey annotation.

Table 5.8a shows that the entire gold standard, including the word forms from manual annotation and annotation from the survey, contains 271 sentences and 250 word forms annotated with a tag *adjective* or *participle*.[25]

Items	**Gold standard**
sentences	222
word forms	250

(a) Total number of sentences and annotated word forms in the gold standard.

Items	**Gold standard**
adjectives	122
participles	128

(b) Total number of annotated adjectives and participles in the gold standard.

Table 5.8b clarifies that the gold standard consists of 122 word forms annotated as adjectives and 128 word forms, annotated as participles, so the proportion of the POS tags in the gold standard was relatively balanced.

3.0.3 Preparation of examples for survey

After disambiguation, some word forms remained ambiguous because they lacked formal syntactic markers of adjectivization. For example, in the phrase *v perevërnutom* mire 'in the inverted

[24] This number excludes three word forms that were annotated as ambiguous in the survey.

[25] *Please Note:* The gold standard excludes three word forms that were tagged as ambiguous in the survey annotation.

world', the word form *perevërnutom* is used without a complement or an adjunct but retains the verbal meaning of the past passive participle 'turned over/inverted'. The word forms such as this one were difficult to assign verbal or adjectival properties using intuition or prior knowledge of Russian grammar. For this reason, I asked a larger group of native speakers to disambiguate these cases. This way, I could avoid bias from my own expectations and rely on native speakers' choices of tags instead. The task was presented to the respondents as a semantic interpretation test in the form of a survey. The purpose of the survey was to obtain tags for the ambiguous word forms selected by a majority of respondents.

As shown in Table 5.9, the survey consisted of 38 sentences, with 39 ambiguous participles and 12 control sentences, which ensured non-randomness of the answers in the survey. Namely, the control examples were used to make sure that respondents carefully considered each example they read and did not make random choices. Control examples contained 12 unambiguous participles that had double adjectival and participial readings given by the analyzer.

Items	**# word forms**	**# sentences**
ambiguous	39	38
controls	12	12
total	51	50

Table 5.9: Overview of ambiguous and control examples.

Both ambiguous and control examples relied on the balanced proportions of syntactic, semantic, and morphological properties. In the ambiguous examples, none of the participles were used with complements, 36 word forms were preposed and two postposed[26] to a head-noun, only nine participles were used with an adverbial modifier, and two were used with adjuncts. Control examples had an equal number of adjectival and participial tags, six participles used with complements and adjuncts of place and six adjectives in stand-alone use with qualitative meanings. Syntactic parameters for control adjectives included adverbs of measure/degree, absence of complements, and use in a preposed position. The semantic component of control participles conveyed the process (movement) and result (physical action, object displacement, and creation) of an action. The semantic component of control adjectives conveyed the quality/property of the defined object. All of the control adjectives were used figuratively and optionally checked as part of collocations using the CoCoCo tool.[27]

Figure 5.3 illustrates the distributions of morphological types of participles in ambiguous and control examples. Present active and past passive participles had the most frequent occurrences

[26] One of the postposed word forms was used predicatively, namely *okazalis′* 'turned out' in *okazalis′ prosty i cinično obnaženy* 'turned out to be simple and cynically exposed'. The other postposed word form was used in an appositive construction *populjacij, kuda bolee prisposoblennyx* 'populations much more adapted'. Although the predicative use does not count as a primary factor of adjectivization, I retained the predicative word form *okazalis′* as an outlier in the survey.

[27] The tool is based on Taiga, RNC and I-Ru corpora, available at: http://cococo.cosyco.ru/index.html

(19 and 17, respectively). Past active and present passive participles had the lowest numbers (10 and 5, respectively). The uneven distribution of morphological types of participles was due to their general distribution in the SynTagRus sample. The great number of sentences retrieved from SynTagRus contained past passive and present active participles, whereas the number of past active and present passive ambiguous participles was much lower.

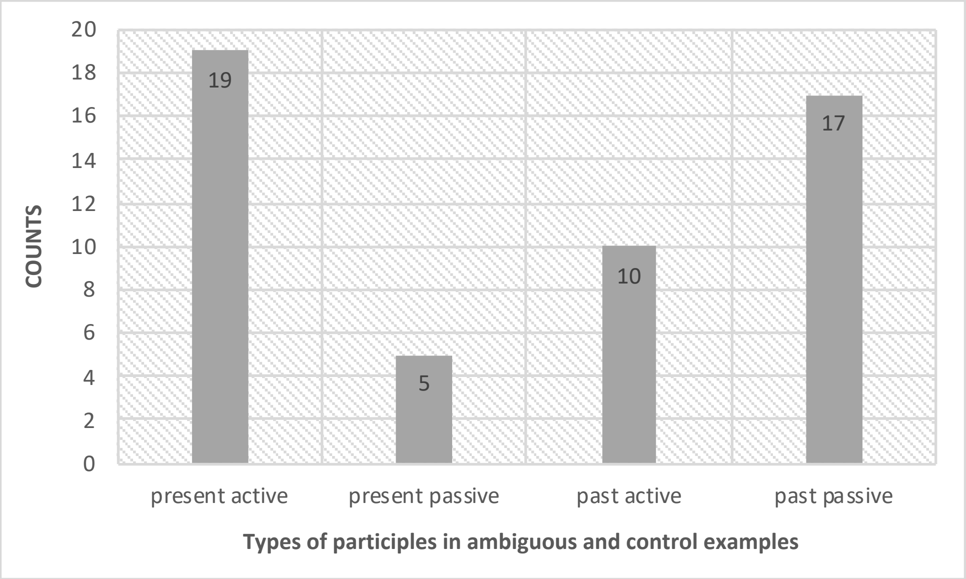


Figure 5.3: Counts of morphological types of participles in the ambiguous and control examples.

3.0.4 Survey design

I constructed a survey in the form of a questionnaire using the *Nettskjema*[28] platform. After compiling ambiguous and control examples, I slightly reduced the context around two ambiguous word forms to make them easier to read. When filling up the survey, I randomized the ordering of control and ambiguous examples. In addition, I mixed the examples that contained different morphological types of participles or different POSs so that there were no repetitive sequences of examples containing one morphological type of participle or an adjective.

The questionnaire consisted of three logical parts. The introductory part explained the purpose of the questionnaire, gave further instructions for filling out the survey, and asked for information on a respondent's native language, gender, and age. The main part consisted of 50 sentences, with the answer field containing three answer variants in the form of a drop-down list. The field for comments on the sentences and the survey was given at the end of the questionnaire.[29] Each sentence had an answer box with three options, including the following:

1. A relative clause, such as *Igrok, kotoryj opozdal* 'The player who was late';

[28] Nettskjema is a tool for designing and conducting online surveys. It was developed and is operated by the University Information Technology Center (USIT), at the University of Oslo. Available at: https://nettskjema.no

[29] The summary report of the survey is presented in Appendix H.

2. A synonymous, unambiguous adjective that conveys a meaning similar to that of the ambiguous participle such as *Nepunktual′nyj igrok* ‘Unpunctual player’;
3. “I cannot say”.

A respondent was expected to select between the first two options, depending on the context conveying adjectival or verbal properties of a word form. The first option implicitly described the word form as an verb, and the second option described it as an adjective. If the respondent hesitated in selecting between the first two options, he or she might select the third option, “I cannot say”. The choice between the options was motivated by respondents’ preferences for selecting answers according to their intuition. If, for some reason, their intuition or knowledge did not help them select an answer, they would choose “I cannot say” to mark their uncertainty. In addition, this third option implied that the tag would be “ambiguous”.

3.0.5 Interpretation of survey results

Forty-three respondents, all native speakers of Russian, gave responses to the survey. Among them, more than half were female (24), and less than half (19) were male respondents. The largest number of the respondents were 30–39 year old adults (19 people), followed by those 50–59 (nine people) and 60–69 (six people) years old.

The percentage of each survey response[30] was recorded. All of the “I cannot say” responses accounted for 10–30% of the total responses. Based on these percentages, I annotated the ambiguous word forms with adjectival (*adj*) or participial (*ptcp*) tags when the following was true:

- 70%–100% of respondents selected adjectival or verbal interpretation;
- 30%–70% of respondents selected adjectival or verbal interpretation, and 0%–30% of respondents selected *adj*, *ptcp* or *ambig*;
- 10%–30% of respondents selected “I cannot say”.

I assigned the tag “ambiguous” or *ambig* to only the word forms for which the difference between Answers 1 and 2 was less than 10%. The rest of the word forms received the tag of an adjective or of a participle.

Figure 5.4 shows that the majority of responses were given equal numbers of participial and adjectival tags (18). There were only three cases in which the difference between the responses for verbal and adjectival readings was minimal, hence the ambiguous tags they received. The relatively even number of votes for verbal and adjectival readings implies that the respondents had a relatively clear understanding of what an adjectival or a verbal interpretation was.

[30] See Appendix 2, Table H.1.

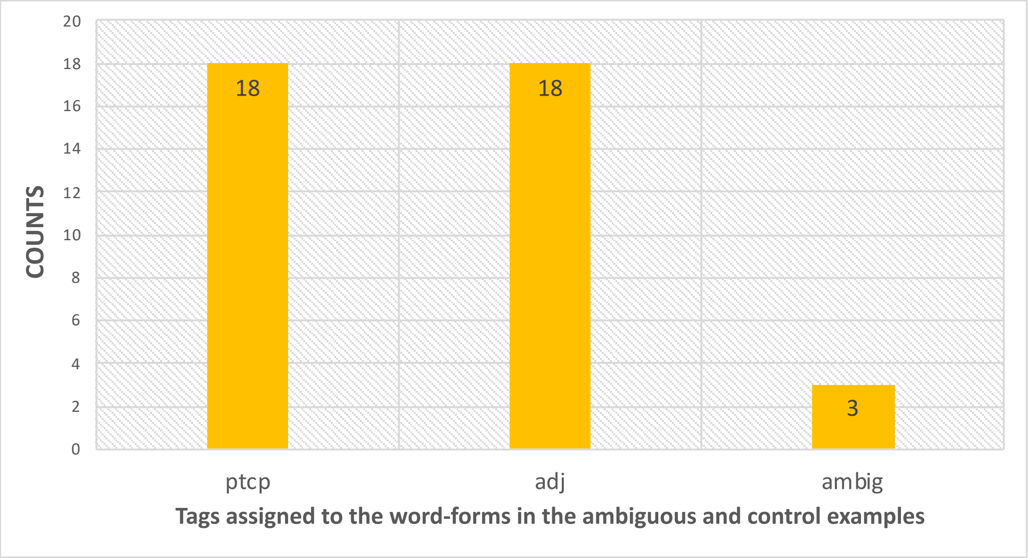


Figure 5.4: The counts of tags (participle as *ptcp*, adjective as *adj* and ambiguous as *ambig*) assigned to the respondents' responses for the ambiguous examples.

The ambiguous sentences are given in (1). In (1a), the word form *suščestvujuščix* 'existing-PRESP/ADJ' is used alone, in a preposed position and not as part of the collocation. In addition, this word form represents a present active participle derived from the verb of existence. Although the syntactic context clearly indicates the absence of verbal properties, the respondents did not show a significant preference for either a verbal or an adjectival reading. The word form *vzaimodejstvujuščie* 'interacting-PRESP/ADJ' in (1b) is also used without complements or adjuncts, preposed to the head-noun *ljudi* 'people' and has the form of a present active participle. Although the analyzer assigned only participial readings to this word form, the respondents split their responses equally between adjectival and verbal readings. Finally, in (1c) the word form *zakrytymi* 'closed-PP/ADJ' is used with the adverbial modifier of time *večno* 'always' but without any complements or adjuncts. This word form represents a past passive participle derived from the verb of impact.

(1) a. *Kogda abiturient govorit s prepodavatelem s glazu na glaz, nikakoj naušnik, daže "samaja malen'kaja iz* ***suščestvujuščix garnitur****", ne spaset.*

When a school-leaver talks to a teacher in person, no headphones, even "the smallest of the **existent headsets**" will not save him. (Sentence 22, survey)

b. *V našix delax očen' važno, čtoby* ***vzaimodejstvujuščie ljudi*** *govorili na odnom jazyke i vsë drug pro druga ponimali.*

'In our business it is very important that **interacting people** speak the common language and understand each other.' (Sentence 26, survey)

c. *Vdol' ulic, produvaemyx tem že vetrom, stojali sutulye doma s večno* ***zakrytymi stavnjami****.*

'Alongside the streets blown through by the wind, there were hunched houses with always **closed shutters**.’ (Sentence 30, survey)

Table 5.10 shows the overall distribution of tags for both ambiguous and control examples after survey completion. Half of the word forms were annotated as adjectives and half as participles (24 word forms each). All of the respondents gave an expected verbal or adjectival interpretation to the control sentences. Only three ambiguous word forms remained after the completion of the survey, implying that the Russian speakers were certain in their choices and could distinguish between adjectival and verbal contexts implied by the answer options.

Word forms	**Survey annotation**
adjectives	24
participles	24
remained ambiguous	3

Table 5.10: The number of participial, adjectival and remaining ambiguous word forms after the respondents’ annotation.

Sentences 22, 25 and 30 contained the ambiguous word forms *suščestvujuščix* ‘existing’, *vzaimodejstvujuščie* ‘interacting’ and *zakrytymi* ‘closed’, which were removed from the gold standard. As the aim of the disambiguation model was to distinguish between adjectival and participial readings, the “ambiguous” tag was not needed in these settings. The word form *obnaženy* ‘exposed’ in Sentence 34 of the survey, voted by over 60% of respondents as adjectival, was eventually tagged as a participle. This helped in complying with the criterion of predicative use that implies verbal use of participles.

3.0.6 Summary

This section shows the construction of the gold standard, small in size but fully adjusted for handling the disambiguation of one type of the POS ambiguity. The sentences extracted from the running texts of the SynTagRus corpus were randomized, but otherwise remained intact at the moment of annotation.

Furthermore, 80% of the corpus is allocated to the participles and adjectives I annotated. These word forms were tagged as ambiguous by the morphological analyzer but used in the clear-cut syntactic contexts and with unambiguous meaning so that I was certain about my annotation choices. In addition, 15% of the corpus consists of the most problematic cases, and 5% of control cases, all of which were annotated by the Russian respondents in the survey experiment. The overall proportion of adjectives and participles in the gold standard is balanced (122 adjectives and 128 participles).

The setup of the survey annotation was based on the morphosyntactic and semantic parameters of the ambiguous word forms in the survey sentences. The parameters were adjusted to ensure that the distribution of control adjectives and participles and morphological types of control participles is balanced. Semantic properties of the control adjectives and participles were also supervised.

The survey allowed me to obtain tags of adjectives and participles by the percentage of respondents' answers regarding adjectival or verbal interpretation. Three word forms received almost equal percentages of votes for adjectival and verbal interpretation, so they were annotated as ambiguous and discarded from the final version of the gold standard. These word forms were omitted to comply with the requirements for the gold standard; that is, to include only adjectival or participial tag, without any remaining ambiguities.

4 Disambiguation experiment

In this section, I discuss the disambiguation experiment, which consists of four main steps. The first step concerns the design of the experiment, ordering and describing CG rules using the factors of adjectivization and weights. In the next step, I define the set of models based on CG rules with syntactic context, morphological properties and weights on one hand, and the statistical baseline model on the other. In the third step, I run the disambiguation and measure its performance. In the last step, I interpret the results from the disambiguation.

The CG models are based on the components that consist of (a) morphosyntactic CG rules, (b) morphosyntactic CG rules combined with the CG weight rule, and (c) the final CG weighted rule to only "select the reading with the highest weight". The baseline model falls into the machine-learning paradigm: It is based on statistical tagging methods of the UDPipe library. The comparison of the CG models with the baseline Google Stanford Dependencies (GSD) model will show how the CG rule-based models can handle the task of the POS ambiguity compared with an existing, easily available probabilistic system. It is also important to cross-check the performance of the CG models in order to identify which rules (syntactic, morphological, or weighted) are most beneficial for disambiguation. Moreover, special focus is given to the performance of the CG rules with weights as compared with the unweighted CG and GSD models, which could highlight the effect of unigram probabilities in solving a morphosyntactic problem.

4.1 Disambiguation models

The distinction of the CG rules with regard to syntactic and morphological constraints, and inclusion or non-inclusion of the numerical matches <*W=MIN*> or <*W=MAX*> is the basic point for defining CG-based disambiguation models. Table 5.11 outlines the disambiguation models used in this experiment.

The CG models include single models CTX and MAX, and extended CTX+MAX, CTX+$\mathrm{W_{MC}}$ +$\mathrm{W_{M}}$, CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX and $\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX models. The single CTX and MAX models use unweighted syntactic rules and one weighted rule ("select the reading with the highest weight"), respectively. The extended models combine unweighted syntactic rules with weighted morphological and syntactic components $\mathrm{W_M}$ and $\mathrm{W_{MC}}$; $\mathrm{W_M}$ consists of weighted morphological rules and $\mathrm{W_{MC}}$, along with weighted morphological and several weighted syntactic rules. Three extended models combine the CTX and MAX models, with or without the weighted morphological ($\mathrm{W_M}$) or weighted morphological and syntactic components ($\mathrm{W_{MC}}$). To make the evaluation of the CG models meaningful, I introduced a baseline model GSD built using machine-learning methods in the UDPipe library[31](Straka et al., 2016). The outcomes of all of these models were compared against the gold standard, consisting of 250 word forms annotated as adjectives and participles.

Model type	Model name
baseline	GSD
CG	CTX MAX CTX+MAX CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$ CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX $\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX

Table 5.11: Disambiguation of CG models defined on the basis of syntactic context and weights. $\mathrm{W_M}$ and $\mathrm{W_{MC}}$ are components added to the CTX and MAX models. GSD is a baseline model trained on the Russian UD treebank using UDPipe.

The GSD model[32] is used as a baseline model trained on the Russian Universal Dependency (UD) treebank *russian-gsd*,[33] annotated and converted by Google. The UD treebank contains tagged/parsed texts from Russian Wikipedia (5030 sentences and 98000 tokens).[34] The UD layout is based on the Google Universal POS tagset (Petrov et al., 2012), the Interset interlingua of morphosyntactic features (Zeman, 2008), and Stanford Dependencies (Tsarfaty, 2013; de Marneffe et al., 2014). To perform the POS tagging and lemmatization, UDPipe uses a MorphoDiTa POS tagger (Straková et al., 2014) implemented as a supervised, averaged perceptron[35] (Collins, 2002), and the classification features adopted from a study by Spoustová et al. (2009). In addition, Figure 5.5 shows an example of a sentence from the gold standard annotated by the GSD model.

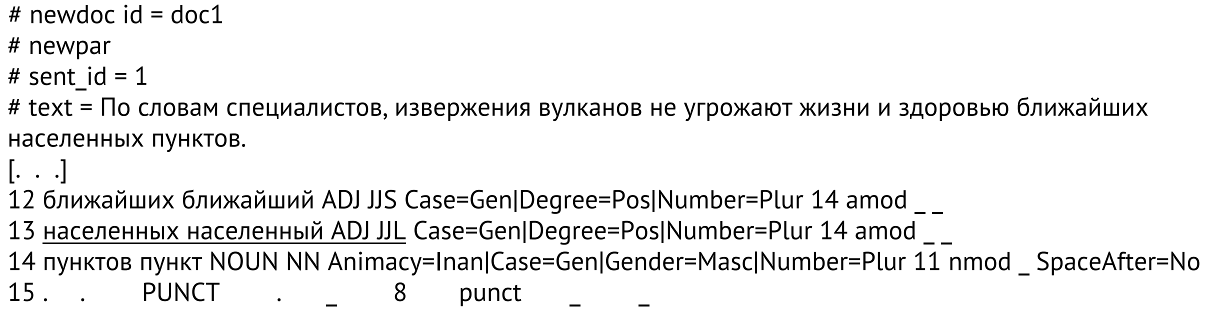
[31] UDPipe is a trainable pipeline for the tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files, available at: https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html

[32] The model (*russian-gsd-ud*-2.4-190531.*udpipe* file) and the UDPipe library are available in the UDPipe R package: https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html#Pre-trained_models.

[33] License CC BY-SA 4.0, available at: https://universaldependencies.org/treebanks/ru_gsd/index.html

[34] Available at: https://universaldependencies.org/treebanks/ru-comparison.html

[35] Perceptron is a learning algorithm that was originally used for training a binary classifier. It represents the simple implementation of a neural network. It finds the best output under the current weights and adjusts weights (that is, parameters) when there is an error. It only updates weights when mistakes are found to ensure that the new weights will be superior to the previous ones in terms of correcting prediction locally.



```
# newdoc id = doc1
# newpar
# sent_id = 1
# text = По словам специалистов, извержения вулканов не угрожают жизни и здоровью ближайших
населенных пунктов.
[. . .]
12 ближайших ближайший ADJ JJS Case=Gen|Degree=Pos|Number=Plur 14 amod _ _
13 населенных населенный ADJ JJL Case=Gen|Degree=Pos|Number=Plur 14 amod _ _
14 пунктов пункт NOUN NN Animacy=Inan|Case=Gen|Gender=Masc|Number=Plur 11 nmod _ SpaceAfter=No
15 .   .      PUNCT     .    _      8      punct     _     _
```

Figure 5.5: An example of the sentence and a word form annotated by the GSD model. The underlined, unambiguous word form *населенных* /naselennyx/'service', as in *населенных пунктов* 'residential areas, settlements' is annotated as a participle with the tags *ADJ JJL.*

This example consists of a document identification number (*newdoc id = doc1*), a sentence ID (*sent_id = 1*), and a plain text of the sentence, followed by each word form, numbered, lemmatized, and annotated. The annotation for the ambiguous word form *населенных* includes the base form *населенный* 'residential, inhabited', the universal POS tag *ADJ*, the language specific tag *JJL*,[36] a set of morphological features for the case (*Gen*) the number (*Plur*), and the universal dependency tag *amod* (adjectival modifier). The ambiguous word form *населенных* is annotated as an adjective *ADJ* and is part of the fixed expression *населенных пунктов* 'settlements'.

The $\mathrm{W_M}$ and **$\mathrm{W_{MC}}$ components** contain weighted morphological rules and some weighted morphological rules that also describe immediate syntactic context. These components were written as a supplement to the extended models and are intended to improve the performance of these models in comparison with the CTX, MAX, and baseline models. The $\mathrm{W_M}$ component is based on the annotation by the analyzer and disambiguation using the weighted rules with morphological properties only. The rules of this component do not specify any syntactic context surrounding the ambiguous word form. In addition, the $\mathrm{W_{MC}}$ component is also weighted and describes general syntactic context only or morphological and syntactic properties combined together. The $\mathrm{W_M}$ rules specify morphological properties, and $\mathrm{W_{MC}}$ specifies morphological properties and/or syntactic context to the right and to the left of the ambiguous word form.

The CTX model is based on the annotation by the analyzer and disambiguation with the CG rules that include constraints on syntactic context (CTX) around the ambiguous word form and are not weighted. The only exception is one rule that specifies the predicative use of the ambiguous word form and does not expand on the surrounding context.[37]

The MAX model is based on the annotation by the analyzer and disambiguation using the MAX rule only. This is a weighted rule that selects only a reading with the highest weight value among participial and adjectival readings (*SELECT:maxweight or MAX*). The MAX rule represents the global weight; it can function independently or be an addition to the $\mathrm{W_{MC}}$ or $\mathrm{W_M}$ components and the CTX models.

[36] This tag stands for an adjective.

[37] See Section 4.2.5.1.

The extended models are expected to perform better because they define more syntactic, morphological factors and weights:

- **The CTX+MAX model** is based on the annotation of the analyzer and disambiguation using the rules from the CTX model and the weighted MAX rule from the MAX model.
- **The CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$ model** is based on the annotation by the analyzer and disambiguation using the rules from the CTX, $\mathrm{W_{MC}}$ and $\mathrm{W_{M}}$ models. This model contains both unweighted/weighted syntactic contexts, and morphological properties.
- **The CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX model** is based on the annotation by the analyzer and disambiguation using the rules from the CTX and MAX models, as well as the $\mathrm{W_{MC}}$ and $\mathrm{W_{M}}$ components. The model uses the same rules as the CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$ model, with an addition of the MAX rule used at the end of the entire rule set.
- **The $\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX model** includes only the annotation by the analyzer and disambiguation using the weighted rules with morphological properties, syntactic contexts, and the MAX rule. Each rule in this model is weighted.

I am primarily interested in the performance of the CTX model, as opposed to the $\mathrm{W_{MC}}$+$\mathrm{W_{M}}$ +MAX model or the CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX model, focusing on distinctions such as weighted and unweighted rules, syntactic context in CTX, and morphological properties in $\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+ MAX. It is also important to see how the MAX model resolves ambiguity alone or with the other models, such as the $\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX and the CTX+$\mathrm{W_{MC}}$+$\mathrm{W_{M}}$+MAX models. Finally, the comparison of the CG models with the baseline GSD model can indicate whether hand-written syntactic and morphological rules and weights, are useful for achieving the best accuracy in comparison with the statistical approach represented by the GSD model.

All of the CG models and their components rely on the syntactic and morphological factors defined in the framework of adjectivization. The CTX model and the $\mathrm{W_{MC}}$ component specify syntactic contexts. Both the $\mathrm{W_{MC}}$ and $\mathrm{W_{M}}$ components contain rules that describe morphological properties that either favor or disfavor adjectivization. Weights of participial and adjectival lemmas used in the MAX model and the $\mathrm{W_{MC}}$ and $\mathrm{W_{M}}$ components reflect the lemma frequencies in Sharoff's frequency list.

4.2 Design and description of CG rules

4.2.1 Overview of CG rules

In this section, I present the design of constraint grammar rules based on the factors of adjectivization discussed in the syntactic and morphosemantic approaches. The weighted morphological analyzer annotates word forms and disambiguates them using CG rules defined in the Russian

CG.[38] After each word form is given morphological analysis, the *vislcg3* parser applies the CG rules defined in the disambiguation models.

I wrote 144 CG rules[39] using information on syntactic and morphological of properties of ambiguous and adjectivized participles, as well as the weights of participial and adjectival lemmas. The syntactic properties also include the description of immediate context (adverbial modifiers, adjuncts, verbal complements, and so forth). The rules are classified into weighted and unweighted rules and ordered accordingly. All of the unweighted rules were written to capture syntactic context. Weighted rules describe syntactic context and morphological properties of participles and specify the weight values of their morphological readings. The rules that specify the syntactic context of an ambiguous word form are also classified according to the number and detail level of contexts they use (that is, specific versus general context). The rules for specific context are designed on the basis of immediate and remote context found in the examples from the test corpora (Section 4.2.2).

Figure 5.6 illustrates the percent distribution of weighted and unweighted rules used for disambiguating ambiguous participles in the Russian CG. More than half of all these rules are unweighted (63%), and less than half are weighted (37%).

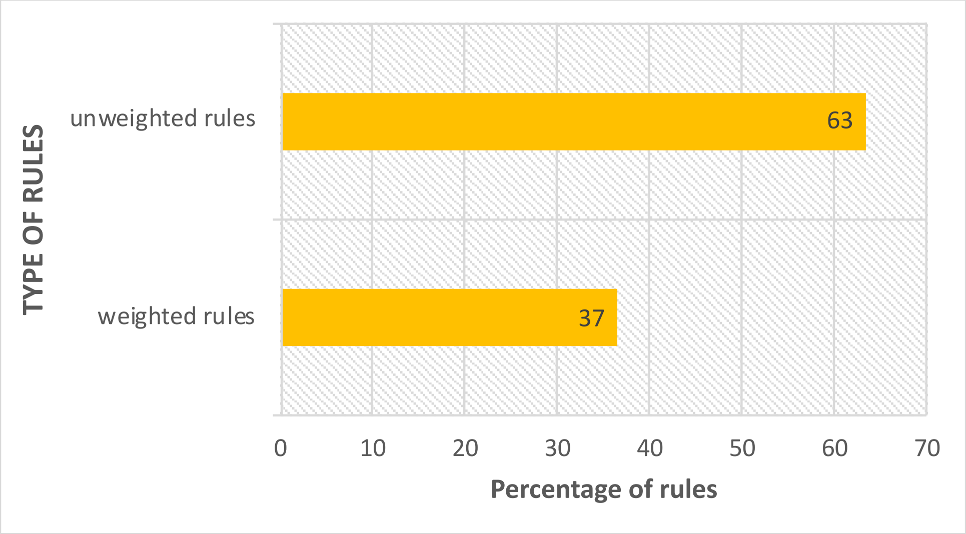


Figure 5.6: Percent distribution of unweighted and weighted rules written for disambiguating participles in the Russian CG.

Weighted rules define a condition for a weight value of a participial or adjectival word form (see Section 2.2). This condition determines whether an ambiguous word form should have a highest or lowest weight for its adjectival or participial readings using the respective numerical matches <*W=MAX*> or <*W=MIN*>. If <*W=MAX*> holds true, the weighted rule selects or removes the the reading(s) of the morphological analysis with the highest weight; if <*W=MIN*> is true, the rule selects or removes the reading(s) with the lowest weight.

[38] The Russian CG is defined in *disambiguator.cg3*, see Chapter 2, Section 4.2.

[39] The entire set of rules, to which I also added the rule *PTCP-GenC-A1* that was written by Francis Tyers and Robert Reynolds, therefore consists of 145 rules.

Figure 5.7 shows that the percentages of the rules with specific and general contexts are almost equal (43% versus 40%), whereas the percentage of the rules that do not describe any context around an ambiguous word form (that is, the rules with morphological properties and MAX rule) is only 17%.

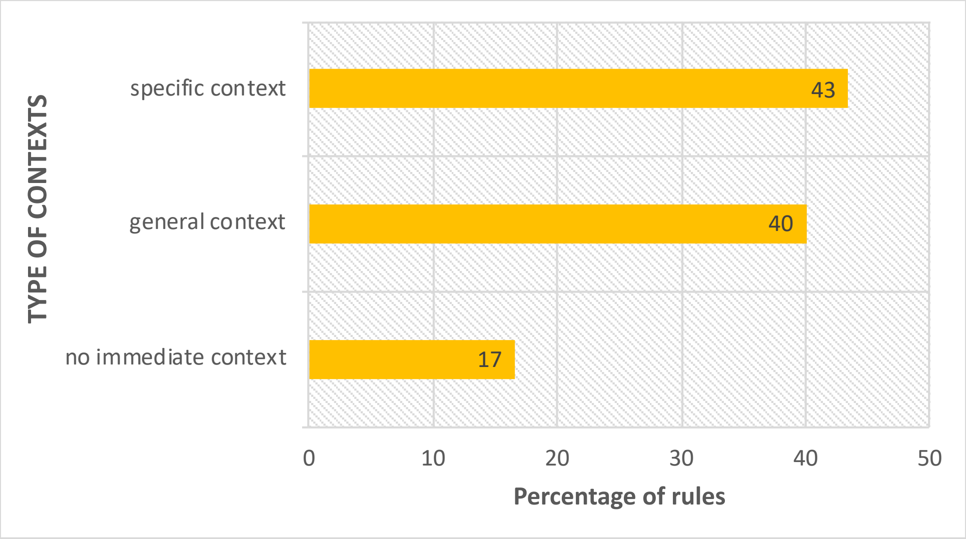


Figure 5.7: Percent distribution of rules with specific and general contexts, and also rules that do not describe any context around an ambiguous word form.

Table 5.12 presents functions of rules used for disambiguating participles.[40]

Functions	**Notation**
remove verbal readings	*REMOVE*:*PTCP-. . . -V*, *REMOVE*:*WPTCP-. . . -V*
remove adjectival readings	*REMOVE*:*PTCP-. . . -A*, *REMOVE*:*WPTCP-. . . -A*
select verbal readings	*SELECT*:*PTCP-. . . -V*, *SELECT*:*WPTCP-. . . -V*
select adjectival readings	*SELECT*:*PTCP-. . . -A*, *SELECT*:*WPTCP-. . . -A*

Table 5.12: Operations and notation of rules for specific contexts. Weighted rules are noted as *WPTCP* while unweighted rules, as *PTCP*.

The rules *REMOVE*:*PTCP. . . V* encode characteristics of syntactic context (such as stand-alone use, adverbs of measure/degree, adverbs of comparative degree, superlative adjectives, and so on) in which adjectivized participles are used. On the other hand, the rules *REMOVE*:*PTCP. . . A* contain constraints describing the context for unambiguous participles (for example, complements, adjuncts, predicative use, and so on). The same principle applies to *SELECT* rules, although they select only supposedly correct readings and remove the rest. For this reason, they can produce more problematic output since they do not disambiguate step by step, as *REMOVE* rules do. There is a greater number of *REMOVE* rules both for adjectival and participial readings because they ensure safer disambiguation in terms of avoiding the selection of incorrect readings.

[40] A set of seven rules used for disambiguating nominalized participles/adjectives was not subject to my analysis in the dissertation. I wrote these rules to avoid further complications in performance caused by the rules for ambiguous participles.

Table 5.13 gives an overview of the number of unweighted and weighted context rules. Unweighted syntactic factors are used in more than half of all the rules (64%), followed by weighted morphological factors (21%) and weighted syntactic factors (14%). The MAX rule representing corpus frequencies, excluding syntactic and morphological factors, amounts to only 1%. Taken together, the syntactic factors (weighted and unweighted) constitute 78% of all the CG rules for distinguishing adjectives and participles, and are likely to have a positive effect on the overall performance of the CTX-based models, compared to the models based on morphological and some syntactic factors (W_{MC}+W_M+MAX).

Order	**Rules**	**Factors**	**#**	**%**
1	unweighted	syntactic factors	93	64
2	weighted	syntactic factors	21	14
3	weighted	morphological factors	30	21
4	weighted	MAX rule	1	1

Table 5.13: Weighted and unweighted rules describing syntactic and morphological factors. Numbers are given in raw counts and percentages.

Notably, I first had to apply the context rules that disambiguate cases when the analyzer outputs triple readings (adjectival, verbal and nominal). The first set includes unweighted rules, followed by the set of weighted rules. The ordering of these rules depends on the density of their context(s). The rules based on morphological properties of participles favoring and disfavoring adjectivization[41] come as the second-to-last set of rules (before the final MAX rule). They test how morphological properties of participles could be relevant in disambiguating ambiguous word forms without falling back on syntactic context.

Table 5.14 illustrates the syntactic and morphological factors of adjectivization included in weighted/unweighted rules. The sequence of adjectives (including a sequence of synonymic adjectives) and participial suffixes that favor/disfavor adjectivization are not asserted as primary factors of adjectivization. Testing the rules using these factors showed that they were recurrent in the test corpora and therefore relevant for the disambiguation experiment.

[41] See Chapter 3, Table 3.2.

Syntactic factors	**Morphological factors**
• verbal complements (direct/indirect objects) • adverbs of measure and degree • adverbs of comparative degree, superlative adjectives • temporal/spatial modification (adverbs, adjuncts) • preposed/postposed position • predicative use • stand-alone use (that is, no adjuncts, complements, adverbial modifiers) • use in a sequence of adjectives	• tense • voice • transitivity • aspect • suffixes (e.g., *-nn/t-*)

Table 5.14: Factors of adjectivization defined in the CG rules used for disambiguating participles, both in specific and general contexts.

The following sections discuss each type of these rules and the contexts that they describe. In addition, the sections focus on the syntactic and morphological factors of adjectivization defined in these rules. To consult all of the rules written for disambiguation of participles, see Appendix I, Table I.1.

4.2.2 Test corpora

I used texts from three corpora to write the rules for specific context and to test and correct all of the CG rules: the RNC, OpenCorpora,[42] and the Yandex-1M corpus.[43] Sentences from the RNC and OpenCorpora were used for defining specific context in the unweighted rules. For example, I looked for an ambiguous word form in the corpora, and as I found one, I looked at the immediate and remote contexts around it and described these in a CG rule. Sentences from the RNC, OpenCorpora, and the Yandex-1M corpus were also used to add more variation to the description of general context in the weighted and unweighted rules.

Each rule (weighted or unweighted) in the CG models was tested on examples from the corpora mentioned above. I selected 1–5 separate sentences with ambiguous word forms from this corpora and ran each rule until the rule output the expected reading. I also tested the rules directly on the texts from OpenCopora and the Yandex-1M corpus (up to 20 sentences per rule), which allowed me to improve the performance of the rules in various syntactic contexts found in the corpora. Furthermore, this helped me to revise and eventually eliminate the most problematic rules or to order the rules in sequence.

[42] Available at: http://opencorpora.org/

[43] Available at: https://translate.yandex.ru/corpus?lang=en

4.2.3 Unweighted CG rules

Unweighted rules specify morphosyntactic context around an ambiguous word form and are classified into the rules for specific and general context. Rules for specific context precede rules for general context because they are more cautious in selecting or removing a reading.

4.2.3.1 Specific context

Rules for specific context describe fine-grained contexts that reflect the syntactic factors of adjectivization. Specific context in the rules has been designed on the basis of sentences from the test corpora. First, specific context encodes various syntactic structures across these sentences and may contain many irregularities, such as in the order of constituents, elliptical constructions, the sequence of adjectives, or adjectives postposed to a head noun. In addition, the length of the context described in a CG rule may vary and can encompass either a sentence part or an entire sentence. Therefore, the rules give a more detailed description of each context with various lengths and constituents.

The example in (2) illustrates the rule *REMOVE*:*PTCP-SpecC-V*10. This rule is designed to disambiguate a word form used in a sequence of adjectives:

(2)
```
REMOVE:PTCP-SpecC-V10 V IF (0 Ptcp) (0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0
N)(-1 CC OR Comm LINK -1 A)(1 Comm OR CC) ;
```

The rule removes verbal readings of an ambiguous word form that:

- is a participle or an adjective agreeing in the same case as a participle (0 *Ptcp*) (0 *A* + *$$NGDAIP*);
- is neither an adverb nor a noun (*NOT* 0 *Adv*)(*NOT* 0 *N*);
- is preceded by a comma or a coordinating conjunction and an adjective (-1 *CC OR Comm LINK* -1 *A*);
- is followed by a comma or a coordinating conjunction (1 *Comm OR CC*).

When applied to the sentence in (3), the rule identifies *počtennyj* 'honorable' as an ambiguous word form, followed by a comma and the adjective *nemolodoj* 'elderly' and removes verbal readings.

(3) *narod nemolodoj,* ***počtennyj****,* *v bol'šinstve semejnyj*
people elderly honor:PP/ADJ mostly married
'people [that are] elderly, honorable, mostly married' (RNC)

4.2.3.2 General context

Rules for general context specify more regular, coarse-grained, immediate context on the basis of the factors of adjectivization. Most of these rules were written on the basis of these factors first, and then tested on the sentences from the test corpora.

The example given in (4) illustrates the rule *SELECT*:*PTCP-GenC-V*2. This rule describes the context with an instrumental complement (without a specified grammatical category).

(4)

```
SELECT:PTCP-GenC-V2 V IF (0 Ptcp)(0 A)(0C NOINS)(NOT 0 Adv)(NOT 0 N)(1
Ins) ;
```

The rule selects verbal readings if an ambiguous word form:

- is a participle or an adjective in any but the instrumental case,[44] not an adverb or a noun (0 *Ptcp*)(0 *A*)(0*C NOINS*)(*NOT* 0 *Adv*)(*NOT* 0 *N*);
- is followed by any word in the instrumental case (1 *Ins*).

For the sentence in (5), the rule detects the ambiguous word form *ispugannyj* 'frightened', its complement (an instrumental noun) *slovami* '[by] words' and selects a verbal reading.

(5) *sprašivaet Dmitriev, vse bolee i bolee udivlennyj i* ***ispugannyj***
asking Dmitriev all more and more surprise:PP/ADJ and frighten:PP/ADJ
slovami soldata
words.INS.PL soldier
'Dmitriev is asking, being more and more surprised and frightened by the soldier's words'
(RNC)

4.2.4 Weighted CG rules

Weighted rules use weights of morphological readings as a primary or an additional constraint. Beyond the obligatory condition for the weight value (maximal or minimal weight), the rules define general syntactic context and morphological properties. However, the MAX rule defines only the weight condition for selecting the highest or lowest weight.

Weighted rules represent the ways morphosyntactic and frequency features can fit the non-contextual disambiguation of adjectivized participles. These rules identify a word form and disambiguate it on the basis of comparing weights in participial and adjectival readings. All of the rules are ordered sequentially so that the disambiguation option of the morphological analyzer removes first verbal and then adjectival readings if the rules identify ambiguous word forms in specific contexts. After that, the *vislcg3* parser selects adjectival and then verbal readings if the rules identify ambiguous word forms in general contexts. The rules for general context were designed on the basis of the syntactic and morphological factors first and then backed up by examples (and their variations) taken from the RNC and OpenCorpora.

[44] Other cases (such as locative, accusative, and so on) were also used in these rules.

4.2.4.1 Examples of weighted rules

Rule *REMOVE*:*WPTCP-V*1.1 which defines a general syntactic context and weight condition is given in (6).

(6)
```
REMOVE:WPTCP-V1.1 V IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Pred)(NOT 0
Adv)(NOT 0 N)(-1 MeasureDegreeAdv)(1 N OR Pron LINK 1 N);
```

The rule removes verbal readings if an ambiguous word form:

- is an adjective with the maximal weight (0 *A* + (<*W=MAX*>));
- is a participle, not used predicatively, not an adverb, not a noun (0 *Ptcp*)(*NOT* 0 *Pred*)(*NOT* 0 *Adv*)(*NOT* 0 *N*);
- is preceded by an adverb of measure and degree (-1 *MeasureDegreeAdv*); and
- is followed by a noun or a pronoun followed by a noun.

For the sentence in (7), the rule detects the word form *nastorožennaja* as ambiguous and removes its verbal reading, as it is preceded by the adverb *večno* and followed by the nouns *naprjažennost′* and *čeloveka*.

(7) *večno* ***nastorožennaja*** *naprjažennost′* *čeloveka*
always alert:PP/ADJ tension person
'always alert tension of a person' (RNC)

The example in (8) shows rule *REMOVE*:*WPTCP-A*7, which specifies a morphological property of the transitive use (TV) and the maximal weight for a participial reading.

(8)
```
REMOVE:WPTCP-A7 A IF (0 A)(0 Ptcp + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)
;
```

More specifically, this rule removes adjectival readings if an ambiguous word form:

- is an adjective (0 *A*); or
- is a participle derived from a transitive verb, with the maximal weight (0 *Ptcp* + *TV* + (<*W=MAX*>)).

The rule *REMOVE*:*WPTCP-A*7 does not search for context because its constraints describe word-internal properties only. For the ambiguous word form *расширенного* /rassirennogo/ 'extended', the rule removes adjectival readings (marking them with ";"). However, it keeps verbal readings, as they contain properties stated in the constraints of the rules (see the cohort below): *PstPss* (past passive participle), *TV* (transitive use), and <W:4.074218750> (the highest weight compared to <W:3.256835938> for adjectival readings). The disambiguated readings for *расширенного* 'extended' are given in Figure 5.8.

```
"<расширенного>"
"расширить" V Perf TV PstPss Msc AnIn Sg Gen <W:4.074218750>
"расширить" V Perf TV PstPss Msc Anim Sg Acc <W:4.074218750>
"расширить" V Perf TV PstPss Neu AnIn Sg Gen <W:4.074218750>
; "расширенный" A Msc AnIn Sg Gen <W:3.256835938> REMOVE:1868:WPTCP-A7
; "расширенный" A Msc Anim Sg Acc <W:3.256835938> REMOVE:1868:WPTCP-A7
; "расширенный" A Neu AnIn Sg Gen <W:3.256835938> REMOVE:1868:WPTCP-A7
:
```

Figure 5.8: The cohort of morphological readings for the output of the word form *расширенного* 'extended' after disambiguation using the rule *REMOVE:WPTCP-A7*.

The MAX rule selects the highest weight (between only the adjectival and participial readings) and applies itself after all the other rules above have been executed. The MAX rule presented in (9) selects the readings with the maximal weight stated in the condition (*<W=MAX>*) if an ambiguous word form is an adjective or a participle (0 *A*)(0 *Ptcp*).

(9) `SELECT:maxweight (<W=MAX>) IF (0 A)(0 Ptcp);`

Figure 5.9 illustrates morphological readings with weights for the word form *населенных* /naselennyx/ 'residential' as in *населенных пунктов* /naselennyx punktov/ 'residential places'. The weighted rule *SELECT:maxweight* (*<W=MAX>*) selects the adjectival reading with the highest weight value 4.15917688, that is, *населённый* /naselënnyj/ 'residential', whereas the participial readings with the verbal lemma *населить* 'inhabit' are discarded.

```
"<населенных>"
        "населённый" A MFN AnIn Pl Loc <W:4.159179688> SELECT:1877:maxweight
        "населённый" A MFN AnIn Pl Gen <W:4.159179688> SELECT:1877:maxweight
;       "населить" V Perf TV Der Der/PstPss A MFN Anim Pl Acc <W:3.637695312> REMOVE:947
;       "населить" V Perf TV PstPss Lxc MFN Anim Pl Acc <W:3.637695312> REMOVE:947
;       "населить" V Perf TV PstPss MFN Anim Pl Acc <W:3.637695312> REMOVE:947
;       "населённый" A MFN Anim Pl Acc <W:4.159179688> REMOVE:947
;       "населить" V Perf TV PstPss MFN AnIn Pl Gen <W:3.637695312> SELECT:1877:maxweight
;       "населить" V Perf TV PstPss MFN AnIn Pl Loc <W:3.637695312> SELECT:1877:maxweight
;       "населить" V Perf TV PstPss Lxc MFN AnIn Pl Gen <W:3.637695312> SELECT:1877:maxweight
;       "населить" V Perf TV PstPss Lxc MFN AnIn Pl Loc <W:3.637695312> SELECT:1877:maxweight
:
```

Figure 5.9: The cohort of morphological readings for the word form *населенных* 'residential' output after disambiguation using the rule *SELECT:maxweight*. The adjectival readings selected by the rule are underlined.

4.2.5 Factors of adjectivization and weights in CG rules

This section discusses linguistic information used for describing context (syntactic) and word-internal (morphological) properties that also include weights. This distinction allows the differentiation to a certain extent of disambiguation models and the observation of how syntactic, morphological information and weights can affect accuracy and correctness of these models' performance. The sequence of these rules was not set before the disambiguation experiment; it was built in progression as each rule was tested on a set of examples from the corpora. It was also easier to begin with the description of the context according to the factors of adjectivization. This is why I first wrote the rules that remove verbal readings, starting from *REMOVE*:*PTCP*. . . *V* rules for specific context.

4.2.5.1 Syntactic factors

The CG rules based on syntactic context are part of the CTX models, and are the first set of rules that were applied to disambiguate the text from the evaluation corpus. These rules, 64% of which are unweighted and 14% are weighted, describe the context, which indicates whether a word form is adjectivized or not, in compliance with the syntactic factors of adjectivization.[45]

Table 5.15 summarizes the set of syntactic characteristics described by the CG rules for specific context. The rules are ordered as the characteristics appear in the table. Rules *REMOVE*:*PTCP*. . . *V* remove participial, and rules *REMOVE*:*PTCP*. . . *A* remove adjectival readings. The syntactic context for an ambiguous word form is described by one or more characteristics because the rules were designed on the basis of corpus sentences.

Operations	**Characteristics of specific context**
REMOVE:*PTCP*. . . *V*	• a preposed/postposed ambiguous word form used with other characteristics • stand-alone use of an ambiguous word form used with other characteristics • an ambiguous word form used with an adverb of measure/degree, with an adjunct as a prepositional phrase • an ambiguous word form used with an adverb of comparative degree or a superlative adjective • an ambiguous word form used with adverbs of time/place • a preposed/postposed ambiguous word form in a sequence of adjectives (e.g., an ambiguous participle preceded by a comma and an adjective) • an ambiguous word form as part of a noun phrase (such as a noun, followed by an ambiguous word form, followed by a genitive noun)

Table 5.15 – continued on the next page

[45] See Chapter 3, Section 4.1.

REMOVE:PTCP... A	• a postposed ambiguous word form used with other characteristics • an short form of an ambiguous participle used predicatively • an ambiguous word form used as a copular verb object • an ambiguous word form used with an adjunct as a prepositional phrase • an ambiguous word form used with a complement (e.g., an agentive instrumental/dative complement as a prepositional phrase) • an ambiguous word form used in a prepositional phrase • an ambiguous word form used with adverbs of measure and degree combined with an agentive instrumental complement • an ambiguous word form used as a prepositional phrase object, with an adverb of time/place • a preposed ambiguous word form used with an adjunct as a prepositional phrase • a preposed ambiguous word form used with an agentive dative complement

Table 5.15: Syntactic characteristics for unweighted CG rules that describe specific context.

A typical example is when an ambiguous participle is preposed or postposed to a head noun and is also used in a sequence of adjectives. Adjuncts are specified according to adjuncts found in the test corpora – for example, an ambiguous word form may be modified by an adverb of measure/degree and followed by an adjunct as a prepositional phrase. A very specific case is the use of an ambiguous participle as an object of the copular verb *быть* /byt′/ ‘be’.

Table 5.16 outlines the set of syntactic characteristics described by the CG rules for general context. The rule *REMOVE:WPTCP... A* (at the bottom of the table) that has no right- or left-side context describes the predicative use of an ambiguous word form. The rules removing adjectival and selecting verbal readings (*REMOVE:PTCP... A*, *SELECT:PTCP... V*) specify the contexts with adjuncts, adverbs of time/place, complements, predicative use, and others. The rules (also weighted ones) selecting adjectival and removing verbal readings (*SELECT:PTCP... A*, *REMOVE:WPTCP... V*, *SELECT:WPTCP... A*) describe ambiguous participles that are used singly, in a sequence of adjectives (in a preposed or postposed position), used as objects of verbs, prepositions, noun phrases, and so forth.

Operations	**Characteristics of general context**
REMOVE:PTCP... A	• an ambiguous word form used with an adjunct as a prepositional phrase • an ambiguous word form used with an adverb of time/place • a short form of an ambiguous word form used predicatively

Table 5.16 – continued on the next page

SELECT:*PTCP*. . . *A*	• an ambiguous word form used with an adverb of time/place • an ambiguous word form used in a sequence of adjectives in a preposed position • an ambiguous word form used in a sequence of adjectives in a postposed position • stand-alone use of an ambiguous word form in a preposed position • stand-alone use of an ambiguous word form in collocations • an ambiguous word form used in a prepositional phrase
SELECT:*PTCP*. . . *V*	• an ambiguous word form used with reflexive pronouns • an ambiguous word form used with an instrumental complement • an ambiguous word form used with a complement as an accusative noun/pronoun • stand-alone use of an ambiguous word form in a postposed position with a dative complement • a postposed ambiguous word form with a nominative/accusative complement • an ambiguous word form used with an adjunct as a prepositional phrase • predicative use of an ambiguous word form • an ambiguous word form used with an adverb of time/place
SELECT:*PTCP*. . . *A*	• non-predicative use of an ambiguous word form with adverbs of measure/degree
REMOVE:*WPTCP*. . . *V*	• an ambiguous word form used with adverbs of measure/degree
SELECT:*WPTCP*. . . *A*	• stand-alone use of an ambiguous word form as a verb/preposition/noun phrase object • stand-alone use of an ambiguous word form • stand-alone use of an ambiguous word form at the beginning of the sentence as a subject • use of an ambiguous word form in a sequence of adjectives
SELECT:*PTCP*. . . *A*	• use of an ambiguous word form in a sequence of adjectives
SELECT:*PTCP*. . . *A*	• predicative use of an ambiguous word form • negation particle *ne* ‘not’ used before an ambiguous word form

Table 5.16: Syntactic characteristics for unweighted and weighted CG rules that describe general context.

The rules that select adjectival or remove participial readings normally use the contexts based on the syntactic factors of adjectivization (atypical for participles), such as no proximity of complements or adjuncts (including adverbs of time and place) or no predicative use. The rules that select participial or remove adjectival readings describe more syntactic context around an ambiguous word form and rely on typical syntactic properties of an unambiguous participle.

4.2.5.2 Morphological factors

The rules based on morphological properties of participles favoring or disfavoring adjectivization[46] come as the second-to-last set of rules (before the final MAX rule). They test how morphological properties of participles could be relevant in disambiguating ambiguous word forms without falling back on syntactic context (in most cases). The morphological rules account for 21% of all rules and are used as components of the CTX and MAX rules. Some of these rules also specify extra conditions, such as the absence or presence of verb complement and the statement of a sentence's end, thus ensuring more safety for removing adjectival or verbal readings.

Table 5.17 illustrates the order of the rule execution, the output of disambiguation (adjective or participle), and morphological properties encoded in each rule. Rules *REMOVE:WPTCP. . . V* disambiguate adjectives, rules *REMOVE:WPTCP. . . A* address participles. Most of these rules specify word-internal properties, with a few rules that also specify an immediate context (that is, passive present imperfective participles followed by a complement). There are no *SELECT* rules among the morphological rules because they do not describe syntactic context (which secures selection of readings) and thus are more prone to erroneous choices. The rules are ordered in such a way that the first set of rules disambiguates adjectives and then participles using several weighted morphological properties. This is followed by the set of rules disambiguating adjectives and then participles using only one morphological property, in addition to weights. Some of these rules also specify the right-side context, such as presence or absence of a verb complement or the end of a sentence. Most of these rules specify word-internal properties with a few rules that also specify an immediate context (for example, a present passive imperfective form used with a complement).

Operations	Morphological properties of an ambiguous word form
REMOVE:WPTCP. . . V	• a present active intransitive word form + not at the end of a sentence, without complement • a present active imperfective word form + not at the end of a sentence, without complement • a past active word form with suffixes *-vš/š-* • a past active perfective word form • a past active perfective intransitive • a present passive perfective word form • a past passive perfective with a complement • a past passive perfective word form • a past passive perfective with a complement • a past passive with suffixes *-nn/t-*

Table 5.17 – continued on the next page

[46] See Chapter 3, Section 4.2.2.

REMOVE:*WPTCP*. . . *A*	• a past passive word form • a present active word form with the suffix *-sja-* • a present active transitive word form • a present active perfective word form with the suffix *-nu-* • a past active imperfective word form • a present passive imperfective word form with a complement • a present passive transitive word form with a complement • a present passive imperfective transitive word form with a complement • a present passive word form with a complement as an instrumental noun/pronoun • a present passive word form with suffixes *-yj/myj-* • a present passive imperfective word form • a present passive transitive word form • a present passive imperfective transitive word form • a present passive word form with a complement as an instrumental noun/pronoun
REMOVE:*WPTCP*. . . *V*	• present passive word form with suffixes *-yj/myj-* • passive word form + not at the end of a sentence • present tense word form
REMOVE:*WPTCP*. . . *A*	• perfective aspect • past tense • imperfective aspect • transitive use • active voice

Table 5.17: Set of weighted rules and the morphosyntactic properties that they describe.

4.2.5.3 Final rule

The MAX rule (that is, *SELECT*:*maxweight* (<*W=MAX*>) *IF* (0 *A*)(0 *Ptcp*);) selects the reading (adjectival or participial) with the highest weight value. It is placed at the very end of the set of the CG rules, which disambiguate the participial word forms. It follows the rest of the rules because it has no context with linguistic information that allows disambiguation. If placed before or among rules with linguistic constraints, it is likely to make more erroneous choices and lead to a greater number of incorrectly disambiguated readings. Its erroneous behavior results from the selection of one POS's reading based only on the comparison of weight values. Thus, placing the rule at the very end forces it to disambiguate the cases that were not resolved by the preceding set of rules. Whether the final choice is erroneous or not does not matter as long as the rule selects readings for one POS.

4.3 Summary

In this section, I present a design of the models for the disambiguation experiment. I define six CG disambiguation models, each containing a set of the CG rules that describe syntactic context only (CTX, CTX+MAX); morphological properties (W_{MC}+W_M+MAX); or both (CTX+W_{MC}+W_M and CTX+W_{MC}+W_M+MAX). Beyond this, 63% of the rules are weighted (using W_{MC} and W_M components), and 37% are unweighted, apart from the stand-alone model MAX that uses only one weighted rule ("select the reading with the highest weight"). The rule is straightforward and unsafe; that is, it does not specify other conditions that would ensure the selection of the appropriate reading is verified enough. For the same reason, the CG parser runs this rule after the rest of the rules. The MAX model is also used jointly with the CTX model and the W_{MC}+W_M components. In addition, 43% and 40% of rules in the CG models describe specific and general context, respectively, whereas only 17% of the rules have no immediate context, as they specify morphological properties. The rules target a wide range of syntactic factors defined by default and were written using various contexts from the test corpora. The rules that deal with morphological factors are based mainly on the morphosemantic approach. A statistical baseline model, GSD, was trained using the UDPipe library, including a neural POS tagger. The comparison of evaluation results of the CG models with the baseline will underline how well the syntactic and morphological information and weights resolve the ambiguity of participles, given an available solution from the machine-learning paradigm.

5 Evaluation

In this section, I review the evaluation metrics for measuring the performance of the models, and discuss their scores for each model. The goal was to define and apply quantitative measures of the effectiveness of the disambiguation models compared to a gold-standard disambiguation. Evaluation can be based on the counts of (a) one tag per word form (adjective or participle or ambiguous), or (b) all tags in the cohort per word form. I chose option (a) because I intended to compare the CG parser[47] to the parsers that always produce one outcome, and to work with examples from the corpus in which only readings for one POS were expected. Option (a) excludes ambiguous readings as a possible outcome of the disambiguation, and satisfies the "one word form–one tag" condition. Choosing option (a) also complied with more specific concerns, as follows:

- The gold standard does not contain ambiguous tags, and the disambiguation models do not have the option of allowing deliberately ambiguous readings.
- The effect of the syntactic, morphological properties, and corpus frequencies used in the CG rules aims to identify and classify a word form as an adjective or as a participle.

[47] The analyzer and the disambiguator.

- Using the “one word form–one tag” correspondence for evaluation makes the scores more transparent and easy to interpret, since each tag from the disambiguation model will be compared to the tag of the corresponding word form in the gold standard.

The cohort-based approach (b) takes the remaining ambiguity as an additional criterion for evaluation into account; on the other hand, it may not fully represent the accuracy of the models in terms of their complete disambiguation, and how weights and linguistic information manage ambiguity.

5.1 Evaluation metrics

The model evaluation consisted of calculating the performance of each model using evaluation metrics. The metrics were based on the distribution of disambiguation outcomes that represented the predictions of an actual (positive/negative) class, and could be true or false. Table 5.18 illustrates a confusion matrix that summarizes the model’s performance with four possible outcomes.

		Actual class (observation)	
		positive class	***negative class***
Predicted class (expectation)	***positive class***	True Positive (TP) correct result	False Positive (FP) unexpected result
	negative class	False Negative (FN) missing result	True Negative (TN) correct absence of result

Table 5.18: Confusion matrix with classification metrics.

In a binary classification, *positive class* represents a given category in which we are interested (such as adjectives), while the remainder of the categories are *negative class* (for example, participles, adverbs, and the like). *False negative* refers to an instance of the positive class being identified incorrectly as negative (false). *False positive* refers to an instance of the negative class being identified incorrectly as positive (false). *True positive* refers to an instance of the positive class being identified correctly as positive (true). *False positive* refers to an instance of the negative class being identified incorrectly as positive (false). *True negative* refers to an instance of the negative class being identified correctly as negative (true).

In this dissertation, the disambiguation models manage two classes, namely adjectives (Task 1) and participles (Task 2), which are word forms that are annotated manually with a single tag in the gold standard. In both tasks, the actual class represents adjectives and participles in the gold standard. In Task 1:

- The predicted positive class represents adjectives tagged by the disambiguation model, and the predicted negative class represents non-adjectives tagged by the disambiguation model.

- The actual positive class represents adjectives in the gold standard, and the actual negative class represents non-adjectives in the gold standard.

The outcomes of Task 1 are as follows:

- TP refers to word forms tagged as adjectives in the gold standard and by the disambiguation model.
- TN refers to word forms tagged as non-adjectives in the gold standard and by the disambiguation model.
- FP refers to word forms tagged as adjectives by the disambiguation model, and as non-adjectives in the gold standard.
- FN refers to word forms tagged as non-adjectives by the disambiguation model, and as adjectives in the gold standard.

The outcomes of Task 2 are as follows:

- TP refers to word forms tagged as participles in the gold standard and by the disambiguation model.
- TN refers to word forms tagged as non-participles in the gold standard and by the disambiguation model.
- FP refers to word forms tagged as participles by the disambiguation model, and as non-participles in the gold standard.
- FN refers to word forms tagged as non-participles by the disambiguation model, and as participles in the gold standard.

The classification metrics used for evaluating the models are precision, recall, F-score, and accuracy. These are standard metrics for binary classification based on the number of outcomes (one outcome per word form; Brownlee 2020a,b). **Precision** quantifies the number of positive class predictions that belong to the positive class. High precision implies a high proportion of TP, and low precision implies a higher proportion of FP. Precision shows how accurate a model is with regard to the predicted positives, and how many of them are actual positives. Precision is calculated using the following equation:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.6}$$

Recall quantifies the number of correct positive class predictions with regard to all positive cases in the dataset. Low recall indicates fewer predictions of the total positive cases. High recall means more predictions of the total positive cases. Recall is the ability of a model to find all the positive cases within a dataset. Recall is calculated using the following equation:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.7}$$

F-score is the harmonic mean of precision and recall expressed in the following equation:

$$\text{F}_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5.8}$$

Harmonic mean allows one to avoid extreme values. It combines the scores of precision and recall in one averaged value.

Accuracy is the proportion of the total number of correct predictions and the total number of predictions made for a dataset. Accuracy quantifies the correct predictions, and thus indicates the usefulness of the model. Accuracy is calculated using the following equation:

$$\text{accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{5.9}$$

5.2 Unresolved ambiguity

Loftsson (2007; also referring to van Halteren et al. 2001) presents two standard reported measures of ambiguity; that is, the ratio of ambiguous tokens and the average number of tags assigned to a token (ambiguity rate). These measures explain the tagging accuracy of the disambiguation models.

The ambiguity rate is used when a tagger does not perform a full disambiguation. It is expressed as the average number of tags assigned to each word form (# stands for number):

$$\text{rate} = \frac{\text{total}\#\text{tags}}{\text{total}\#\text{wordforms}} \tag{5.10}$$

It is preferable that the ambiguity rate is low, and is not greater than 1.

The ratio of ambiguous word forms to the total number of word forms is expressed by the following equation:

$$\text{ratio} = \frac{\text{ambigWordforms}}{\text{total}\#\text{wordforms}} \tag{5.11}$$

Figures 5.10 and 5.11 show the ambiguity measures across the models.

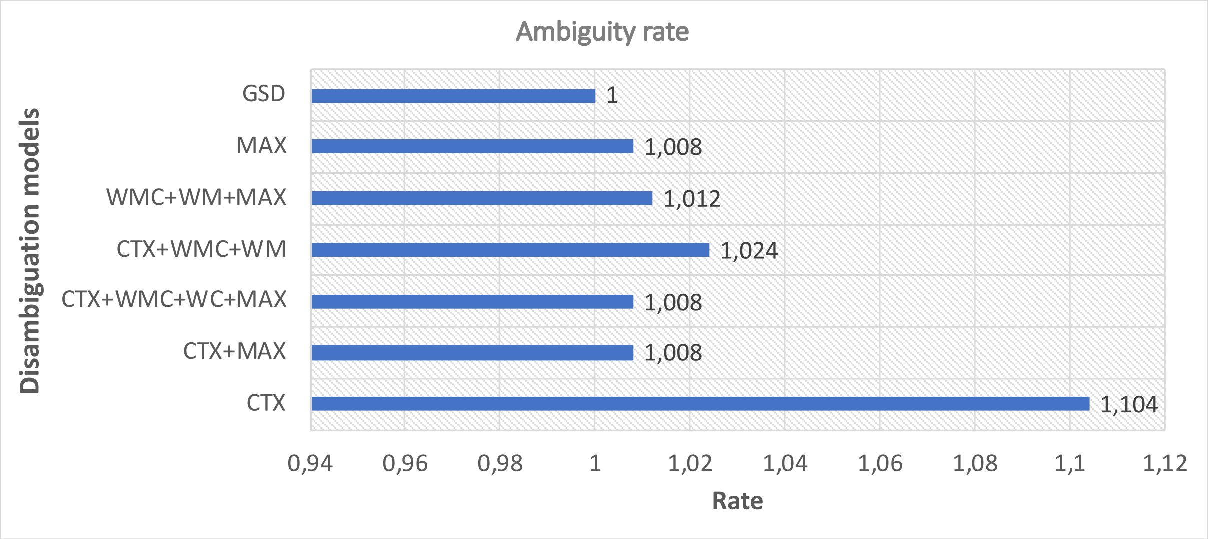


Figure 5.10: Ambiguity rate across the disambiguation models.

Figure 5.10 indicates that the GSD model had an equal number of tags and word forms (that is, one tag per word form) with a rate of 1. Three models containing the MAX rule had a slightly greater number of tags per word form. The lowest rate was that of the CTX model (1.104), which does not have the weighted rules.

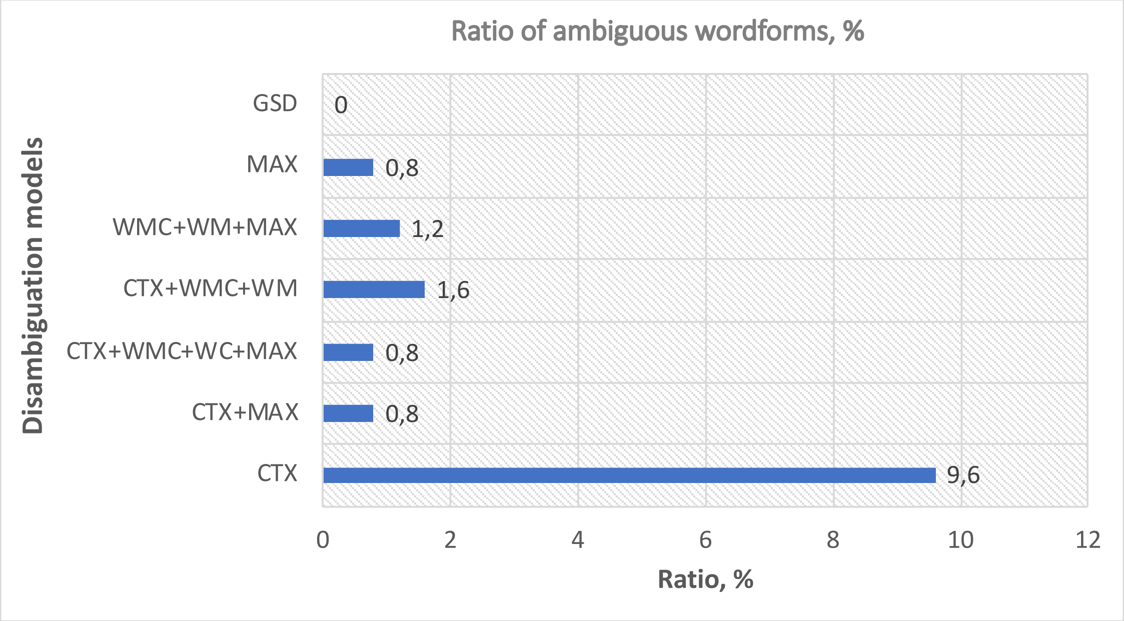


Figure 5.11: Percentage ratio of ambiguous word forms across the disambiguation models.

Figure 5.11 represents the percentage of unresolved ambiguity per model in more detail. The greatest amount of ambiguity was produced by the CTX model (9.6%) without the MAX rule. By contrast, the MAX model and the joint CTX+W_{MC}+W_{M}+MAX models had the second lowest ratio of ambiguity (0.8%). The remaining ambiguity in the CG models is represented by double or triple readings:

- adjective, participle
- participle, particle
- adjective, participle, noun
- adjective, noun
- participle, noun
- participle, particle

The main reasons for unresolved ambiguity were as follows:

- insufficient coverage by the CG rules (particularly the unweighted ones);
- the removal of adjectival or participial readings by the CG rules;[48] and
- the absence of entries in the verbal lexicon of the analyzer (rare cases).

The GSD model had no ambiguous tags (an ambiguity rate of 1), but eight word forms were tagged as nouns, one as a determinant, and one as an infinitive form.

Figure 5.12 illustrates the output of the CTX model for the word form *закрытыми* 'closed'. The readings *A MFN AnIn Pl Ins* <*W*:4.690429688> and *V Perf TV PstPss MFN AnIn Pl Ins* <*W*:4.464843750> remained ambiguous because the rules did not capture the context surrounding this word form, and did not use weights for the comparison.

```
"<закрытыми>"
"закрыть" V Perf TV PstPss Lxc MFN AnIn Pl Ins <W:4.463867188>
"закрытый" A MFN AnIn Pl Ins <W:4.690429688>
"закрыть" V Perf TV PstPss MFN AnIn Pl Ins <W:4.464843750>
;  "закрыть" V Perf TV Der Der/PstPss A MFN AnIn Pl Ins <W:4.464843750> REMOVE:1268:Der
:
```

Figure 5.12: Output of the CTX model wherein the participial and adjectival readings were not disambiguated.

Figure 5.13 shows the output of the MAX model that did not compare weights <*W*:3.375976562> (adjectival reading) and <*W*:3.9218750> (participial reading) because the latter was removed by the rule *REMOVE*:874 preceding the set of rules that disambiguated participles.

```
"<несущими>"
"несущая" N Fem Inan Pl Ins <W:0.0>
"несущий" A MFN AnIn Pl Ins <W:3.375976562>
;        "нести" V Impf IV Der Der/PrsAct A MFN AnIn Pl Ins <W:3.9218750> REMOVE:874
;        "нести" V Impf IV PrsAct MFN AnIn Pl Ins <W:3.9218750> REMOVE:874
:
```

Figure 5.13: Output of the MAX model wherein the participial and adjectival readings were not disambiguated.

An example of an FN for the participial word form *очищенные* 'purified' as in *очищенные водой* 'purified by water' is provided in Figure 5.14. The weighted rule did not work for this

[48] These rules either precede the CG rules in the disambiguation models or constitute the rules in these models.

cohort because adjectival readings were missing. The condition for the contexts *IF* (0 *A*)(0 *Ptcp*) did not hold for this cohort; therefore, the rule did not compare weights and did not disambiguate the readings.

```
"<очищенные>"
"очищенная" N Fem Inan Pl Acc <W:0.0>
"очищенная" N Fem Inan Pl Nom <W:0.0>
"очистить" V Perf TV PstPss MFN AnIn Pl Nom <W:3.7656250>
"очистить" V Perf TV PstPss MFN Inan Pl Acc <W:3.7656250>
; "очистить" V Perf TV Der Der/PstPss A MFN AnIn Pl Nom <W:3.7656250> REMOVE:1268:Der
; "очистить" V Perf TV Der Der/PstPss A MFN Inan Pl Acc <W:3.7656250> REMOVE:1268:Der
:
```

Figure 5.14: Output of the MAX model with missing adjectival readings in the cohort.

In the following sections, I describe the evaluation based on the multi-class classification available via the Python module *sklearn.metrics*.[49] The metrics were based on a comparison of two or more lists of tags. The source list contained tags from the gold standard, and the target list contained tags from the evaluation models. Each tag taken from the gold standard list was compared to its corresponding tag taken from the target list. The comparison of tags from these lists provided the outcomes for each class of tags ('adjective' or 'participle'). These outcomes were then used to compute the scores for the evaluation metrics.

5.2.1 Use of the CG weighted rule in the MAX model

The MAX model demonstrates a considerably different performance from the performances of the CTX models (CTX+MAX, CTX+W_{MC}+W_{M}, CTX+W_{MC}+W_{M}+MAX). This model disambiguates the word forms using only one weighted rule, without any other rules being involved in the process. This rule leads to straightforward sorting into correct and incorrect readings through the *SELECT* operation. The rule selects the reading(s) with the highest weight values, and only considers the context to verify that a word form has both adjectival and participial readings.

Figure 5.15 illustrates the cohort of morphological readings for the word form *открытый* /otkrytyj/ 'open', as in the *открытый смех* 'genuine laughter' output by the morphological analyzer. Each adjectival lemma *открытый* 'open', as in *"открытый"A Msc AnIn Sg Nom*, is assigned the weight of 5.168945312 (<*W*:5.168945312>), and each participial lemma *открыть* 'open', as in *"*открыть*" V Perf TV PstPss Msc AnIn Sg Nom*, received the weight of 4.641601562 (<*W*:4.641601562>).

[49] Available at: https://scikit-learn.org/stable/modules/model_evaluation.html

```
"<открытый>"
"открытый" A Msc AnIn Sg Nom <W:5.168945312>
"открытый" A Msc Inan Sg Acc <W:5.168945312>
"открыть" V Perf TV Der Der/PstPss A Msc AnIn Sg Nom <W:4.641601562>
"открыть" V Perf TV Der Der/PstPss A Msc Inan Sg Acc <W:4.641601562>
"открыть" V Perf TV PstPss Msc AnIn Sg Nom <W:4.641601562>
"открыть" V Perf  TV PstPss Msc Inan Sg Acc <W:4.641601562>
:
```

Figure 5.15: The cohort of the morphological readings with weights for the word-from *открытый* 'open' before disambiguation.

As shown in Figure 5.16, the values for weights were obtained from the verbal and adjectival lexicons, in which each lemma is assigned a specific weight. Following the morphological analysis, the disambiguation weighted rule was applied, and had only one directive, which was to select the reading that was assigned the maximum weight according to the condition *SELECT:maxweight* (*<W=MAX>*).

```
открыть:откр св_12а_ы́ть "weight: 4.642003549666032" ; (verb.lexc)
открытый:откры́т п_а "weight: 5.168373709205478" ; ! Z 1a (adejctives.lexc)
```

Figure 5.16: The cohort of the morphological readings with weights for the word form *открытый* 'open' before disambiguation.

Figure 5.16 illustrates the cohort of morphological readings for an adjective that the MAX rule disambiguated correctly (a true positive). The weighted rule compared the values of all the available weights in the cohort, and then selected the readings with the maximum weights. In this case, it selected the adjectival readings (*A Msc AnIn Sg . . .*) because the weight of the adjectival lemma *открытый* 'open/opened' (5.168945312) was greater than was the weight of the verbal lemma *открыть* 'open' (4.641601562).

```
"<согнутый>"
"согнутый" A Msc AnIn Sg Nom <W:3.392578125> SELECT:1877:maxweight
"согнутый" A Msc Inan Sg Acc <W:3.392578125> SELECT:1877:maxweight
;        "согнуть" V Perf TV Der Der/PstPss A Msc AnIn Sg Nom <W:3.221679688> REMOVE:1268:Der
;        "согнуть" V Perf TV Der Der/PstPss A Msc Inan Sg Acc <W:3.221679688> REMOVE:1268:Der
;        "согнуть" V Perf TV PstPss Msc AnIn Sg Nom <W:3.221679688> SELECT:1877:maxweight
;        "согнуть" V Perf TV PstPss Msc Inan Sg Acc <W:3.221679688> SELECT:1877:maxweight
:
```

Figure 5.17: The cohort of the morphological readings with weights for the word-from *открытый* 'open' before disambiguation.

Figure 5.18 shows the cohort of the morphological readings for the participial word form *согнутый* /sognutyj/ 'bent', as in *[кусочек проволоки] согнутый в локте* '[little piece of wire] bent at the elbow'. In this example, the maximum weight was assigned to the verbal lemma согнутый 'crooked' in the readings *"согнутый" A Msc AnIn Sg . . .* because its weight of

3.392578125 was higher than was the weight of 3.221679688 for the participial lemma *согнуть* /sognut′/ 'bend'. Therefore, the weighted rule selected *V Perf TV . . .* as the best reading and removed all the other readings. This case is a false positive because it was tagged as a participle in the gold standard and as an adjective in the MAX model.

```
"<согнутый>"
"согнутый" A Msc AnIn Sg Nom <W:3.392578125> SELECT:1877:maxweight
"согнутый" A Msc Inan Sg Acc <W:3.392578125> SELECT:1877:maxweight
;       "согнуть" V Perf TV Der Der/PstPss A Msc AnIn Sg Nom <W:3.221679688> REMOVE:1268:Der
;       "согнуть" V Perf TV Der Der/PstPss A Msc Inan Sg Acc <W:3.221679688> REMOVE:1268:Der
;       "согнуть" V Perf TV PstPss Msc AnIn Sg Nom <W:3.221679688> SELECT:1877:maxweight
;       "согнуть" V Perf TV PstPss Msc Inan Sg Acc <W:3.221679688> SELECT:1877:maxweight
:
```

Figure 5.18: The cohort of the morphological readings with weights for the word form *открытый* 'open' before disambiguation.

The examples above show that the weighted MAX rule performs straightforward sorting using numerical values of weights as a basis, and selects the highest value via a comparison of these values. No context is involved in the disambiguation. The MAX model appears to show a greater percentage of false negatives and false positives, probably due to a bold selection of any reading that is greater in weight than another. Nevertheless, since many lemmas from the verbal and adjectival lexicons have weights, most of the readings were disambiguated in a straightforward manner, and little ambiguity remained. There were only two occurrences of remaining ambiguity in the text disambiguated via the MAX rule; one occurrence was due to a missing lemma in the verbal lexicon, and the other to the removal of the verbal reading by a different set of CG rules that are not used for resolving participle-adjective ambiguity. The latter was the result of the removal of participial readings from the morphological analysis of an ambiguous participle using the CG rule *REMOVE*:874 preceding the set of the CG rules that I wrote.

5.3 Evaluation methods

Evaluations of model performance can be done in one of two ways, either by omitting the remaining ambiguity or by retaining it as a separate class. The ambiguity across almost all the models, with the exception of GSD, may have weakened the scores for precision and accuracy. It is also likely to have had a negative impact on recall scores, as it mainly affected the word forms that should have been tagged as adjectives or participles but were not. However, I consider the negative impact to be important for understanding how the CTX and MAX models contributed to minimizing the ambiguity rate, as they also disambiguated based on the highest weight value. For this reason, I suggest two types of evaluation:

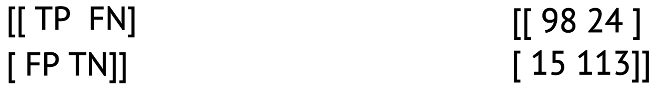
1. **Evaluation 1**: Replacing ambiguous tags with single ones; and
2. **Evaluation 2**: Retaining ambiguous and/or erroneous tags as a third but unwanted class; thus, any word form that remains ambiguous is either FN or TN. For example, in the

evaluation of the ‘adjective’ class, the word form tagged as ambiguous by the model and as an adjective in the gold standard becomes FN, while the one tagged as ambiguous by the model and as a participle in the gold standard becomes TN.

I used Evaluation 1 as the default type, and Evaluation 2 to show the weight performance in the recall scores.

5.3.1 Evaluation 1

Evaluation 1 used a method that consisted of automatically replacing ambiguous tags (treated as Not a Number (NaN) values[50]) with disambiguated tags. Such a technique is used in classification or any other task involving missing values in the dataset. Missing values are replaced by their neighboring values (values to the right or to the left of the missing value). This approximation can make precision and recall more balanced. I used the method *ffill* (forward fill[51]), which propagates the previous value forward; that is, it replaces the missing value with the nearest preceding non-missing value. If an NaN value is preceded by *adj*, it will be replaced by *adj*; if it is preceded by *ptcp*, it will be replaced by *ptcp*. The confusion matrix and the summary report of the metric scores for the model CTX now have two classes (‘adjective’ and ‘participle’). Figure 5.19a illustrates the slots for TP, FN, FP, and TN in a confusion matrix. The confusion matrix for evaluating the classification of adjectives using *ffill* is presented in Figure 5.19b.



```
[[ TP  FN]
 [ FP TN]]
```

```
[[ 98 24 ]
 [ 15 113]]
```

(a) The scheme for a confusion matrix in Evaluation 1.

(b) A confusion matrix for the CTX model in Evaluation 1.

Figure 5.19: Confusion matrix in Evaluation 1, classification of adjectives.

The outcomes for the class ‘adjective’ consist of the following scores:

- 98 adjectives are classified as adjectives (TP).
- 15 participles are classified as adjectives (FP).
- 24 adjectives are classified as participles (FN).
- 113 participles are classified as participles (TN).

Precision and recall were computed using these scores:

[50] NaN is used to represent entries that are undefined.

[51] The *ffill* method is part of the Python library *Pandas*; the documentation on filling missing values using the specified method is available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.fillna.html

- precision (‘adjective’):
  $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) = 98/(98+15) = 0.867$

- recall (‘adjective’):
  $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) = 98/(98+24) = 0.803$

If we want to evaluate the classification of participles, the notation of outcomes in the confusion matrix is reversed, as shown in Figure 5.20.

```
[[ TN  FP]
 [ FN TP]]
```

Figure 5.20: The scheme for a confusion matrix for the classification of participles in Evaluation 1. The corresponding confusion matrix is given in Figure 5.19b.

The outcomes for the class ‘participle’ consist of the following scores:

- 113 adjectives are classified as adjectives (TP).

- 24 participles are classified as adjectives (FP).

- 15 adjectives are classified as participles (FN).

- 98 participles are classified as participles (TN).

The summary report presented in Table 5.19 provides the metric sores for each class, the overall accuracy (0.844), and the average scores for the entire model with 250 word forms (under the *Support* column), including for adjectives (122 word forms) and participles (128 word forms).

	Precision	**Recall**	**F1-score**	**Support**
adjective	0.867	0.803	0.834	122
participle	0.825	0.883	0.853	128
accuracy			0.844	250
macro avg	0.846	0.843	0.843	250
weighted avg	0.846	0.844	0.844	250

Table 5.19: Summary report for the metric scores in the disambiguation experiment for the CTX model, Evaluation 1.

- Global accuracy (*accuracy*) is the proportion of correctly classified tags in the total number of all the tags; for example, accuracy $= (98+113)/250 = 0.844$.

- The macro averaged score *(macro avg)* is the mean of each class’ precision/recall/f1-score; for example, macro avg precision $= (0.867+0.825)/2 = 0.846$.

- Weighted averaged score (*weighted avg*): each class' precision/recall/f1-score weighted by the number of tags for each class; for example, weighted avg precision $= (0.867 * 122 + 0.825 * 128)/250 = 0.846$.

For the further evaluation of the other models, I used **global accuracy** and **weighted average precision**, **recall** and **f1-score** for both classes.

5.3.2 Evaluation 2

In this evaluation, I defined unresolved ambiguity as a separate class ('ambiguous') in addition to the classes of adjectives and participles. This class will not be predicted because it is not found in the gold standard. However, it is used to decrease the recall scores of the other classes. The schematic representation of the confusion matrix is presented in Figure 5.21a. If we want to assess the classification of the class 'adjectives', the confusion matrix is represented by the following outcomes in Figure 5.21b.

```
[[ TP  FN  FN]
 [ FP  TN  TN]
 [ FP  TN TN]]
```

(a) The scheme for the confusion matrix in Evaluation 2.

```
[[ 98   4  20]
 [  0   0   0]
 [  5  20 103]]
```

(b) A confusion matrix for the classification of adjectives in Evaluation 2.

Figure 5.21: Confusion matrix for the CTX model in Evaluation 2.

The outcomes for the class 'adjectives' consist of the following scores:

- 98 adjectives are classified as adjectives (TP).
- 4 adjectives are classified as ambiguous tags (FN).
- 20 adjectives are classified as participles (FN).
- 0 (1st column) ambiguous tags are classified as adjectives (FP).
- 0 (2nd column) ambiguous tags are classified as ambiguous tags (TN).
- 0 (3rd column) ambiguous tags are classified as participles (TN).
- 5 participles are classified as adjectives (FP).
- 20 participles are classified as ambiguous tags (TN).
- 103 participles are classified as participles (TN).

The summary report in Table 5.20 indicates metric scores for each class, the overall accuracy (0.804), and the average scores for the entire model.

	Precision	Recall	F1-score	Support
adjective	0.951	0.803	0.871	122
ambiguous	0.000	0.000	0.000	0
participle	0.837	0.805	0.821	128
accuracy			0.804	250
macro avg	0.596	0.536	0.564	250
weighted avg	0.893	0.804	0.845	250

Table 5.20: Summary report for the metric scores for the CTX model in the disambiguation experiment, Evaluation 2.

5.4 Model performance

Before reporting the metric scores of the models, I rounded them to two decimal places, according to the recommendations for reporting statistical results (Cumming et al., 2012).[52]

Table 5.21 presents the precision and recall scores for "adjective" and "participle" classes. All of the models have higher precision and lower recall scores for adjectives, and lower precision and higher recall for participles. This means that all of the models tended to accurately disambiguate more adjectives than participles but were biased towards selecting participles rather than adjectives. In terms of recall, CTX and MAX showed the greatest difference in scores for adjectives and participles. More explicitly, the recall score in the CTX model was 0.80 for ADJ and 0.88 for PTCP, whereas the recall score in the MAX model was 0.66 for ADJ and 0.73 for PTCP. The least difference between precision and recall for adjectives and participles occurred between the W_{MC}+W_M+MAX and GSD models. CTX+W_{MC}+W_M+MAX had the best f1-scores (0.85 for ADJ and 0.86 for PTCP). The lowest f1-score arose in the MAX model, which was 0.68 for ADJ and 0.71 for PTCP.

#	Model	Precision		Recall		F1-score	
		ADJ	PTCP	ADJ	PTCP	ADJ	PTCP
1	CTX	0.87	0.82	0.80	0.88	0.83	0.85
2	CTX+MAX	0.86	0.84	0.82	0.88	0.84	0.85
3	CTX+W_{MC}+W_M	0.87	0.84	0.82	0.88	0.84	0.86
4	CTX+W_{MC}+W_M+MAX	0.87	0.84	0.83	0.88	0.85	0.86
5	W_{MC}+W_M+MAX	0.71	0.71	0.69	0.73	0.70	0.72
6	MAX	0.70	0.69	0.66	0.73	0.68	0.71
9	GSD	0.79	0.78	0.76	0.80	0.78	0.79

Table 5.21: Performance of the models for classifying adjectives (ADJ) and participles (PTCP), Evaluation 1. Metrics are rounded to two decimal places.

[52] The literature displays that there are no explicit guidelines regarding the number of significant digits for effect sizes (e.g., Cousineau, 2020).

Table 5.22 below summarizes the performance of the disambiguation models expressed in weighted average precision, recall, f-1 score, and global accuracy using Evaluation 1. As the scores were averaged, differences between the scores of the models may have been attenuated, and they now look different from the scores in Table 5.21.

As the remaining ambiguity and the non-participial and non-adjectival tags in the GSD model have been replaced by adjectival/participial tags, the distribution of FP and FN have been smoothed so there is no great difference between precision and recall scores across most of the models. For example, the CTX model had a precision of 0.85 and a recall and accuracy of 0.84. Beyond this, the weighted CTX+MAX and CTX+W_{MC}+W_M models did not differ in their precision and differed only slightly in their recall in comparison to the scores of the CTX model. For example, the precision of CTX+MAX and CTX+W_{MC}+W_M was 0.85, and the recall of these models was 0.85, compared with the recall of 0.84 for CTX.

#	Model	Weighted avg precision	Weighted avg recall	Weighted avg f1-score	Accuracy
1	CTX	0.85	0.84	0.84	0.84
2	CTX+MAX	0.85	0.85	0.85	0.85
3	CTX+W_{MC}+W_M	0.85	0.85	0.85	0.85
4	CTX+W_{MC}+W_M+MAX	0.86	0.86	0.86	0.86
5	W_{MC}+W_M+MAX	0.71	0.71	0.71	0.71
6	MAX	0.70	0.70	0.70	0.70
9	GSD	0.78	0.78	0.78	0.78

Table 5.22: Overall performance of the models, Evaluation 1. Metrics are rounded to two decimal places.

The CTX models, weighted or unweighted, performed considerably better than the GSD and MAX models. The accuracy scores of the CTX models equaled or surpassed 84%, whereas the accuracy for MAX and W_{MC}+W_M+MAX was equal or slightly over 70%. In other terms, the syntactic rules in the CTX models resolved the ambiguity of participles much better than non-context based models and became better if they were combined with the weighted rules and/or morphological component. The baseline model surpassed the weighted non-CTX models by showing the best scores in all metrics (0.78).

The MAX rule slightly improved the results in recall and accuracy at 0.01 in the CTX+MAX model and did the same with the weighted morphological component in CTX+W_{MC}+W_M. Thus, both of these models were accurate at 85%. Using the weighted rules which specify morphological properties and syntactic context in CTX+W_{MC}+W_M and CTX+W_{MC}+W_M+MAX improved precision and recall in these models up to 0.02, as compared with the CTX model with only the unweighted syntactic rules (for example, from 0.85 to 0.86 for precision and from 0.84 to 0.86 for recall).

The non-CTX CG models performed worse compared with the CTX-based models and the baseline model. Without any syntactic and/or morphological rules, the MAX model disambiguated at around 70% accurately. Combining the weighted morphological properties and syntactic contexts with the MAX rule (W_{MC}+W_M+MAX) gave the highest recall and f1-score of 0.71. The CTX models had higher scores than the baseline model, whereas the weighted non-CTX models performed worse than the GSD model and had precision, recall and accuracy below 0.78. The models MAX and W_{MC}+W_M+MAX disambiguated differently (accuracies of 70% and 71%, respectively). When these models were added to CTX (that is, CTX+W_{MC}+W_M, CTX+MAX), they disambiguated much better (accuracy of 85% and above).

Table 5.23 below illustrates the performance scores for the models using Evaluation 2. I resorted to this type of evaluation to understand (a) to what extent a specific model manages to reduce ambiguity when it is needed and (b) how this reduction affects the metric scores of the model. This is why I focused mainly on recall and accuracy scores in my interpretation of the models' performance.

#	Model	Weighted avg precision	Weighted avg recall	Weighted avg f1-score	Accuracy
1	CTX	0.89	0.80	0.85	0.80
2	CTX+MAX	0.85	0.84	0.85	0.84
3	CTX+W_{MC}+W_M	0.85	0.84	0.85	0.84
4	CTX+W_{MC}+W_M+MAX	0.86	0.85	0.86	0.85
5	W_{MC}+W_M+MAX	0.72	0.71	0.71	0.71
6	MAX	0.70	0.69	0.69	0.69
7	GSD	0.79	0.76	0.78	0.76

Table 5.23: Overall performance of the models, Evaluation 2. Metrics are rounded to two decimal places.

As shown in Evaluation 1, the replacement of the ambiguous tags/erroneous tags using the *ffill* method led to relatively equal scores in the CTX-based models. However, one can better grasp the differences among the CTX models that maintain the unresolved ambiguity as a separate class. The low recall score of CTX (0.80) markedly improved when weighted rules were added, and the score increased up to 0.84–0.85 in CTX+MAX, CTX+W_{MC}+W_M and in CTX+W_{MC}+W_M+MAX. Combining the W_{MC}+W_M and MAX rules in the CTX model provided a greater recall score of 0.85 at the cost of lowering precision from 0.89 to 0.86. Adding weighted morphological and syntactic rules and the MAX rule on top of the CTX rules gave the best accuracy score, 0.85. This implies that weights resolve ambiguity at the cost of decreasing precision but improving recall and accuracy of the models that include the CTX rules.

The evaluation scores of GSD were relatively high compared with the non-CTX CG models and were closest to the CTX and W_{MC}+W_M+MAX models in terms of recall and accuracy. For example, the GSD recall score of 0.76 was 0.05 higher than that of W_{MC}+W_M+MAX (0.71) and 0.04 lower than the recall score of CTX (0.80). Among the weighted non-CTX models, W_{MC}+W_M+MAX had the largest precision score of 0.72 and the accuracy score of 0.71. As a comparison, the MAX model showed a precision of 0.70 but a recall and global accuracy of 0.69.

The high recall scores in the weighted CTX models were gained by lowering their corresponding precision scores. The metric scores of the non-CTX models did not differ considerably from their corresponding scores in Table 5.22, so replacing ambiguous tags or keeping them did not affect the performance of these models, as their unresolved ambiguity was already low.[53]

5.5 Summary

In this section, I summarize my observations and interpretations I made in the previous section; these mainly concern Evaluation 1 unless Evaluation 2 is mentioned explicitly.

First, the disambiguation models differed in their amounts of unresolved ambiguity. The baseline GSD model had no ambiguity, the MAX, CTX+MAX and CTX+W_{MC}+W_M+MAX models had only 0.8% ambiguity, and the weighted CTX+W_{MC}+W_M and MAX+W_{MC}+W_M showed 1.6% and 1.2% of ambiguity, respectively. Thus, the unweighted CTX model leaves the highest ratio of ambiguity after disambiguation compared with the weighted CG models or the baseline GSD. Of all the weighted models, the MAX model, particularly the baseline model, showed the lowest ratio for ambiguity. The unresolved ambiguity in the CG models occurred when certain CG rules did not manage to disambiguate the word form, or they erroneously removed adjectival or participial readings and left other readings (such as a participle and a particle, a determinant, or a noun). This could also have occurred due to the absence of an entry in the verbal lexicon of the analyzer.

All of the disambiguation models (both CG and baseline) showed a tendency to tag more adjectives correctly than they did participles; that is, they displayed a high degree of precision for adjectives and low levels of precision for participles. In addition, the models selected more relevant participles than adjectives, as indicated by a high recall for participles and a low recall for adjectives. This bias towards adjectives in terms of precision and towards participles in terms of recall was most noticeable for the CTX model. Its performance showed that, even when many more adjectives were annotated correctly, fewer word forms were selected as relevant adjectives, and more were selected as participles (in other words, high precision and low recall). Furthermore, the CTX model annotated word forms as adjectives correctly with 0.87 precision, whereas it identified relevant adjectives with a recall rate of 0.80 .

Syntactic rules used alone in the CTX model performed quite well both in terms of precision and recall, with 84% accuracy. The MAX rule improved recall and precision slightly when combined with the CTX rules. The rules with the morphological/syntactic constraints (W_{MC}+W_M) used with the CTX rules performed in the disambiguation task more correctly and in a safer manner, thus achieving better precision and recall scores than did the MAX model. Thus, weights combined with morphological properties/syntactic context performed better than did the MAX rule.

The MAX and W_{MC}+W_M rules used together in the CTX model showed more noticeable

[53] See Section 5.2, Figures 5.10 and 5.11.

improvement in the precision and recall scores, with the CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX having the best performance at 86% accuracy. In Evaluation 2, the CTX-based models with the fewest unresolved ambiguities had the best performance; that is, the CTX+WMC+WM+MAX achieved 85% accuracy, and the CTX+MAX and CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$ models both had 84% accuracy.

When the MAX rule was not bound to the context, it disambiguated with an accuracy rate of 70%, but it did not exhibit any significant improvement when combined with the $\mathrm{W_{MC}}$+$\mathrm{W_M}$ components. The average performance of the weighted models without syntactic contexts was comparable to that of the baseline model in terms of low precision and recall, although GSD precision was the highest compared to the precision of the non-CTX weighted models. Thus, the simple weighted disambiguation was similar to the baseline disambiguation of the GSD model, with slightly lower precision and recall, as it fell within the range of 66% and 71%.[54]

Since the ambiguity ratio for the CTX model was not significant (9.6%, Figure 5.11), it is not yet clear whether the weighted models are superior in terms of removing ambiguity after disambiguation. The role of the model using the simple and straightforward weighted MAX rule can be minimized by the set of elaborated rules describing syntactic contexts (as in CTX). However, the weighted models would work more efficiently with a larger amount of remaining ambiguity given a larger corpus.

6 Conclusion

The disambiguation experiment was found to be a complex task, even though it addressed only one type of POS ambiguity. The experiment involved the implementation of weights based on the corpus frequencies in Sharoff et al.'s (2013) frequency list; that is, the transformation of the corpus frequencies of adjectival and participial lemmas into weights, and the annotation of verbal and participial lemmas in the verbal and adjectival lexicons of the morphological analyzer for Russian. In addition, the experiment included the design of the CG rules using syntactic and morphological properties discussed in the framework of adjectivization. The performance of the weighted/unweighted CTX models was superior to the performance of the baseline GSD model and the MAX models (that is, MAX and $\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX). The MAX models with and without additional components remained closer to the baseline GSD model in terms of overall accuracy.

The disambiguation experiment shows that the frequency encoded in the weights is sufficient for resolving the ambiguity of participles. The weights help to eliminate as many ambiguous readings as possible, thus decreasing the amount of unresolved ambiguity to the minimum. Furthermore, the models using weights combined with the syntactic context and morphological properties show better results than do the CG models with weights only.

Weights used alone in the MAX model resolve ambiguity with 70% accuracy, but which word forms fall within the 30% missing from the accuracy score remains unclear. Weights also

[54] See Tables 5.21 and 5.22.

perform slightly better when combined with the syntactic and morphological components in the $\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX model. In terms of the remaining ambiguity, the MAX model improves the recall and accuracy scores when added to the CTX-based models. Weights in the MAX rule are derived from distributions of high- or mid-frequency ranks of lemmas; thus, this rule may be more likely to fail when it disambiguates atypical (low-frequency) cases. The weights function as a valid support for the CTX-based models because they increase the recall scores at the cost of decreasing the precision scores. They also leave almost no unresolved ambiguity, for example, the GSD model. The performance of the weighted models allows for the conclusion that frequency does help to resolve POS ambiguity, but not as efficiently as do context-based rules or the statistical tagger in the GDS model.

The GSD model's performance falls in between the performances of the CTX models and the weighted MAX models. It provides a sufficient disambiguation without leaving any ambiguous tags, and retrieves and tags a proportional number of adjectives and participles correctly. The GSD performs better than does the weighted CG models without the syntactic rules and does not leave any ambiguity, similar to the MAX model. Unlike the CTX models, the GSD assigns few erroneous tags to the ambiguous word forms.

Formalizing the syntactic and morphological factors of adjectivization in the CG rules seems to allow for the differentiation between adjectivized and unambiguous participles in the most accurate way. The rules with syntactic contexts in the CTX models perform well in terms of both precision and recall. The weighted rules specifying morphological properties (as well as the MAX rule) alone do not perform well in the non-CTX models, and are comparable to the baseline model. However, they improve the performance of the CTX +$\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX model in terms of precision and recall. The components with morphological properties and syntactic contexts improve the CTX model to a greater extent than does the MAX rule. This may imply that the use of morphological properties as complementary information for syntactic properties may help to distinguish formally between adjectivized and unambiguous participles.

The performance of the CTX models combined with the weighted components of the MAX models can be improved by increasing the size of the test corpus and adding more CG rules. The survey annotation of some ambiguous cases in the gold standard may have deviated from the choices made by expert annotators or constraints defined in the CTX and weighted rules. This might imply that the intuition of native speakers can contradict some of the formalisms provided in the CTX rules. For example, the word form *ustarevšij* 'outdated', as in *v ustarevšix zakonax* 'in the outdated laws', was considered to be a participle by the majority of the respondents in the survey. However, the weighted CG models, such as CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$, CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX, and MAX, tagged the word form *ustarevšij* as an adjective. In these CTX models, the rule *SELECT*:*WPTCP-A*2.2[55] used in the $\mathrm{W_{MC}}$ component, and the MAX rule in the CTX+MAX and MAX models, selected adjectival readings of the word form. Only the GSD model tagged

[55] The rule selects an adjectival reading if a word form is part of a prepositional phrase without adjuncts or verbal complements.

the word form as a participle, in accordance with the gold standard. The disambiguation models may be improved by (a) using several syntactic rules that capture the specific context and the W_{MC}+W_M/MAX rules for the rest of the contexts, and (b) combining the syntactic rules for general contexts with W_{MC}+W_M/MAX rules.

All of the models tend to select more word forms as participles and to tag adjectives with more accuracy than they did participles. It remains unclear whether this bias is caused by the gold standard rather than by the syntactic context, morphological properties, or the weights of the disambiguated word forms. The examples in the gold standard reflect both typical (which I disambiguated manually) and atypical/problematic (disambiguated by the respondents) uses of word forms as adjectives or as participles. In addition, 20% of the atypical uses might have complicated the disambiguation task. The preference for tagging adjectives as participles by the CG disambiguation models could be explained by the semantics of adjectives, their idiomatic readings, and their stand-alone use without explicit syntactic context. The reason that more participles tend to be disambiguated correctly is due to their overt syntactic context, such as verb complements, adjuncts, or adverbial modifiers (to a lesser degree).

More generally, the performance of the disambiguation models implies that syntax is definitely more significant for recognizing a grammatical category than are the internal properties of a word form, such as morphology and corpus frequency. Although the global weights (represented in the MAX rule) allow the models to disambiguate with 70% accuracy, they appear to fit general uses of participles and adjectives, and make more errors in the atypical cases. Finally, both weights and morphological properties improve the disambiguation only as an addition to the syntactic rules.

Chapter 6

Conclusions

This chapter provides a summary of the dissertation, discusses the implications drawn from the theoretical framework and the analysis of adjectivization, and provides some directions for future research.

In this dissertation, I developed an approach to the adjectivization of participles in Russian that encompassed three domains, namely the theoretical framework, the empirical analysis of factors pertaining to adjectivization, and the development of the disambiguation model. I constructed a theoretical framework to attain the main objectives of the dissertation, which were to identify the factors that underlie adjectivization and to provide a quantitative assessment of their relationship with the ambiguity of participles. In addition, I applied this framework as a practical basis for resolving participle-adjective ambiguity in Russian text. The disambiguation task (a) applied Constraint Grammar (CG) rules based on the factors of adjectivization and the corpus frequencies of adjectives and participles represented by weights, and (b) estimated the effectiveness of these rules for making successful distinctions (according to the gold-standard corpus) between adjectives and participles in the evaluation corpus.

From the theoretical perspective, I investigated adjectivization as a synchronic phenomenon that leads to the POS ambiguity between adjectives and participles observed in the corpus data, and which arises from the affixless change of the syntactic category (that is, conversion) without any morphological changes. Based on the review of the main approaches to adjectivization, the dissertation brought into focus factors that make a formal distinction between an adjectivized and an unambiguous participle on the syntactic level, and account for the development of adjectival and the loss of verbal properties in a participle that becomes adjectivized. With regard to the factors of adjectivization, I analyzed the syntactic behavior of ambiguous participles and their internal, semantic and morphological, properties (such as tense, voice, aspect, transitivity, and lexical meaning) that favor or obstruct adjectivization. The morphosemantic properties also explain the causes and show the results of this phenomenon. The analysis revealed that the syntactic context allowed for the differentiation between adjectivized and unambiguous participles based on the immediate syntactic context surrounding a participial word form, while the internal properties of participles are capable of contributing to adjectivization (positively or negatively), and of predetermining the process of its development.

From the empirical perspective, I explored how specific morphosyntactic factors of adjectivization, such as the adverb of measure and degree *očen′* 'very', and the grammatical categories of tense, aspect, voice, and transitivity, manifested across the corpus data (the Araneum Russicum and the Russian National Corpus, or RNC) by means of corpus frequency distributions. Furthermore, I quantitatively assessed the statistical significance of these factors, together with the effect of the ratio of participles and the ipm (instances per million) frequency of the base verbs, with regard to the ambiguity of participles.

In addition to the quantitative exploratory analysis of adjectivization, I developed a constraint-based disambiguation model that resolves participle-adjective ambiguity. This model was based on formalized factors of adjectivization and the corpus frequencies of participial and adjectival lemmas expressed as weights. The design of the model is novel, as it combines weights and linguistic properties to address the ambiguity of participles. The weights are corpus frequencies that were log-transformed and scaled according to the corpus size; therefore, their interpretation is meaningful in terms of the frequency rank to which a participle or an adjective belongs. The dissertation thus shows that a purely linguistic analysis (based on morphosyntactic properties and lexical frequencies) can be used to construct a method for solving applied tasks related to text processing.

1 Summary

Chapter 1 introduced the main research questions and the objectives of the dissertation. First, I focused on adjectivization as systematic and problematic POS ambiguity observed in the corpus data. I then defined the research questions and the scope of the dissertation, and presented a brief theoretical background to adjectivization as a linguistic phenomenon. Finally, I pointed out the contribution of the research, and provided an outline of each chapter in the dissertation.

The overview of the main tools and methods used in Chapters 3 and 4 was presented in Chapter 2. These methods enabled morphological annotation and subsequent disambiguation using weights and CG rules. They were used in the exploratory analysis of adjectivization and in the development of the disambiguation model that resolves participle-adjective ambiguity. I first provided basic algebraic definitions of transducers and weights, as well as the domains of their applications, including:

- machine translation
- speech recognition
- tagging
- lexical processing
- optical character recognition (OCR)

- text summarization and generation

I then discussed the functions of the morphological analyzer, the lexicons associated with it, and the formalism of CG (Karlsson, 1990). The morphological analyzer for Russian is a weighted, finite-state transducer that annotates text via a morphological analysis, and assigns a weight to each of the word forms in a text. Disambiguation is made possible by the *vislcg3* parser that uses the CG formalism, a surface-oriented and morphology-based parsing. The Russian CG consists of grammatical categories, entity definitions and rules that represent the syntactic or morphological property of a word form, and remote or immediate constituents of the syntactic context surrounding an ambiguous word form.

In Chapter 3, I described the development of the theoretical framework for the adjectivization of participles, departing from the POS homonymy and conversion. I discussed the notions of lexical and syntactic ambiguity, and narrowed these down to a POS homonymy with related morphological forms and meanings. Adjectivization is a type of this POS homonymy that arises from conversion, an affixless change of syntactic function maintaining morphological expression (*cf.* Lieber, 2005; Manova, 2011; Štekauer et al., 2012). I then outlined the factors entailed in adjectivization based on the morphosyntactic and semantic properties of ambiguous (adjectivized) and unambiguous participles. Syntactic factors (Say, 2016; Timberlake, 2004) manifesting themselves in the immediate syntactic context surrounding a participial word form appeared to be the most salient for differentiating between ambiguous and unambiguous word forms. Attributive and predicative functions do not play prominent roles in differentiating between verbal and adjectival uses of participles, although specific predicative uses of participles are likely to imply that a participial lexeme is adjectivized (such as present active participles used predicatively with the copular *byt′* ‘be’, as in *rezul′taty byli udručajuščimi* ‘the results were discouraging’).

The factors of adjectivization discussed in the morphosemantic approach (Kustova, 2012; Kalakuckaja, 1971; Kolochkova, 2011; Černega, 2009) refer to the grammatical and semantic meanings of participles, as well as to the meanings of their base verbs. The meanings expressed by the grammatical categories of tense, voice, aspect, and transitivity have a positive or negative effect on the loss or retention of verbal properties. Abstract or idiomatic meanings of the base verbs favor a semantic shift to the adjectival meaning in their corresponding participles. The gradual extension of the participial meaning to an adjectival one is the primary criterion for semantic change in this approach. Tense and voice are viewed as primary grammatical categories that affect adjectivization; the present tense in present active participles tends to lack temporality, and the passive voice in past passive participles implies a reduced argument structure. For this reason, present active and past passive participles are regarded as the most pervasive group of adjectivized participles, as discussed by Kustova (2012) and Kalakuckaja (1971). Aspect affects adjectivization by interacting with transitivity, voice, and tense. For example, the perfective aspect and the past tense combined together reinforce the resultative verbal meaning, and favor adjectivization in past active participles. Furthermore, the processual (verbal) meaning of the imperfective aspect disfavors adjectivization, while the resultative and atemporal meaning of the perfective aspect favors it. The properties of transitive/intransitive participial forms, such as argument structure and verb government, also affect adjectivization. When a participle is used transitively, it is less likely to be adjectivized because its argument structure enables verb government and the joining of verbal complements. When a participle is used intransitively, its reduced argument structure and the lack of verb government obstruct adjectivization. Thus, semantics and the grammatical properties of participles affect the process of adjectivization positively or negatively, and undergo changes in the course of adjectivization.

In Chapter 4, I provided a quantitative assessment of the significance of several morphosyntactic factors pertaining to adjectivization (that I discussed previously) across two sets of Russian corpora, the Araneum Russicum 1.20 GB corpus and the manually disambiguated version of the RNC. In the first study, I quantitatively and qualitatively assessed the relevance of the syntactic factor, the adverb of measure and degree *očen′*, by comparing the construction *očen′* used with present tense finite verbs to the construction *očen′* used with their corresponding participial forms in the Araneum corpus. The comparisons of the ratios of these two constructions showed that the adverb *očen′* was used with finite verbs more frequently than it was with participles. The semantic classes of participles demonstrated a statistically strong association with their ratios; however, this tendency was more statistically robust for the *očen′* construction used with finite verbal forms than for those with participles. The study revealed that the adverb of degree *očen′* was not an independent factor of adjectivization because it was used easily with finite verbal forms, and with unambiguous and adjectivized participles. Moreover, the intensification enabled by this adverb was initially triggered by the gradable meaning of these verbal forms; thus, the intensification of participles via *očen′* relies equally on their gradable semantic components inherited from their base verbs. Compared to other syntactic criteria of adjectivization, such as a preposed position to a head noun, a lack of complements and temporal/spatial modifications by adjuncts or adverbs, among others, the adverb *očen′* is marginal as a factor in adjectivization, and indicates that a participle is adjectivized if the factors mentioned previously are in place.

In the second study, I explored the relationship of the pervasiveness[1] of participles to the rank-frequency distribution of their corresponding base verbs, as well as to the distributions of their morphosyntactic properties of tense, voice, aspect, and transitivity. I first examined the distributions of participial lemmas based on the ratio of participial to (finite/infinitival) verbal lemmas, ordered by the ipm frequency and the ranks of verbal lemmas. High-frequency verbs tended to form more finite/infinitival forms, while low-frequency verbs appeared to have more participles than they did other verbal forms. Among the high-frequency verbs, I found more adjectivized present active participles; among low-frequency verbs, I found more past passive adjectivized participles.

[1] Pervasiveness is expressed as the ratio of participial to verbal finite/infinitival forms.

I then referred to the features of tense, voice, aspect, and the transitivity of participles to statistically predict their ambiguity based on the output of the morphological analyzer for Russian: double ambiguous readings ADJ/PTCP and unambiguous readings PTCP. The results of the statistical analysis confirmed several claims proposed by Kustova (2012) and Kalakuckaja (1971) in their approach to adjectivization. The ambiguous past passive participles were the most pervasive morphological type (that is, they had the highest ratio); moreover, past passive participles had a strong effect on predicting ambiguous forms of participles. Nevertheless, as a single predictor, the past tense demonstrated a significant effect on unambiguous readings, which may be attributed to the strong markedness of this tense, conveying a more defined temporal meaning and thus obstructing adjectivization. The ambiguous past active participles represented the second most pervasive morphological type of participles, although they were not mentioned in Kustova's (2012) study as the most numerous group of adjectivized participles, and were not significant predictors of ambiguity. Transitive perfective ambiguous participles showed the highest ratio among the other groups of participles. They were presumed to be numerous because they form two types of participles, namely past active and passive participles. A finding that has not been discussed previously in the approaches to adjectivization concerns the effect of frequency on ambiguity. More specifically, the frequency rank of the base verbs and the pervasiveness of participles strongly predict when a participle is ambiguous. As the observations in the scatter plots show (*cf.* Chapter 4, Figures 4.10 and 4.11), top-frequency verbs had more ambiguous than they did unambiguous participles; the remaining high and mid-frequency rank verbs formed fewer participles than they did finite/infinitival forms. The verbs from the low-frequency rank formed more participles, among which there were more ambiguous than unambiguous forms. The significance of the frequency rank of the base verbs and the pervasiveness of participles for ambiguity (that is, ambiguous forms) was also confirmed by the statistical analysis.

The development of the disambiguation model, including the implementation of weights, the design of the gold standard, the disambiguation experiment and its evaluation, was presented in Chapter 5. First, I introduced my approach to weighting and the implementation of weights in the Russian verbal and adjectival lexicons and the morphological analyzer of Russian; as a result, the weights of verbal and adjectival lemmas were output as part of the morphological analysis by the transducer. Weights were also used as part of the CG rules for resolving participle-adjective ambiguity, either alone (for example, the very last, weighted MAX rule) or combined with the rules describing syntactic context and/or morphological properties. Second, I developed a gold-standard corpus by crowd-sourcing the annotations of the most difficult cases of ambiguity from the SynTagRus corpus using a larger group of native Russian speakers. The corpus allowed the resolution of only one type of ambiguity; for this reason, it was relatively small (128 participles and 122 adjectives) in comparison to the evaluation sets commonly used in machine-learning, for example. At the same time, the gold standard was relatively noise-free and had balanced distributions of adjectives and participles in relation to their proportion to each other, as well as a balanced number of morphological types of participles. Accordingly, the overall distribution of the ambiguous word forms in the gold standard was not skewed. Finally, I designed and ran the disambiguation models, and compared them to the baseline Google Stanford Dependencies (GSD) model based on the UDPipe stochastic parsing and pre-trained on the Russian Universal Dependency treebank *russian-gsd*. For the CG disambiguation models, I wrote 144 rules, including weighted or unweighted syntactic rules and morphosyntactic rules, and the final weighted MAX rule ("select the reading with the highest weight"). The sequence of the rules represented the actual order of their execution by the *vislcg3* parser. The syntactic and morphological rules were based on the factors of adjectivization defined in Chapter 3 and on weights. I ran the disambiguation models on the plain text of the gold standard and evaluated their performance, paying particular attention to how the weights and the rules based on the syntactic context performed. All the disambiguation models tended to tag more adjectives than participles correctly, and chose more relevant participles instead of adjectives. The best performance (in terms of weighted averaged metric scores) was shown by the CG model combining syntactic and morphological rules with weighted constraints (CTX+W_{MC}+W_M+MAX), which yielded above 80% accuracy. The weight-only MAX rule when used alone was only accurate at 70%; however, when used as a component in the other CG models based on syntactic context or morphological properties, it improved their metric scores and reduced the remaining ambiguity.

2 Implications

In this dissertation, adjectivization was approached both as a syntactic phenomenon dependent on the internal grammatical and semantic properties of participles, and as a task concerning the POS ambiguity resolution. Since morphosyntactic ambiguity is generally systematic in corpora, finding a linguistic-based solution for disambiguating adjectivized participles and adjectives is an interesting task in itself, and has the potential of being extended to further tasks that resolve POS ambiguity, such as noun-adjective, adverb-preposition, and so on.

Adjectivization is attributed to the ambivalent nature of participles, which allows them to shift from the verbal to the adjectival paradigm. The shift includes semantic changes and changes of grammatical meaning in the morphosyntactic properties of tense, voice, aspect, and transitivity. Adjectivized participles are placed in between unambiguous participles that demonstrate a clearly verbal syntactic behavior and unambiguous deverbal adjectives that used to be participles but now function only as adjectives. Adjectivized participles may partially or fully lose their verbal properties and affinity with the meaning of their base verbs; the likely indicator of the degree (full or partial) of adjectivization is the semantic affinity with the base verb: The looser the connection, the less verbal a participle is, and the more adjectival behavior the participle shows in the syntactic context.

The syntactic behavior of adjectivized participles is not something that appears on its own. The process of change in the lexical semantics and grammatical meanings of the morphosyntactic properties of participles precedes the change in the syntactic function, and the subsequent development of adjectival syntactic behavior. A participle that has morphosyntactic properties that account for the reduction of the argument structure, a lack of temporal meaning, and/or inherits a qualitative/abstract meaning from its base verb, is more inclined towards adjectivization (that is, past passive and present active participles). The greater the verbal meaning and the more verbal properties a participle has, the less it will favor adjectivization.

Thus, the dissertation shows that factors concerning adjectivization related to syntactic context account for the endpoints of the process. The semantic and morphological properties may begin the process, encourage its development or obstruct it; such semantic factors as a lack of affinity with the base verb and the extension of lexical meaning are also the result of adjectivization, and may be triggered by changes in the grammatical meaning of a participle or by its qualitative/abstract meaning derived from the base verb. I analyzed the process of adjectivization synchronically without considering the stages of its development over specific time periods from the diachronical perspective. Further investigation of adjectivization as a diachronic change using a historical corpus of Russian may reveal more tendencies in the syntactic contexts of adjectivized participles. The claims of the morphosemantic approach were confirmed by the joint corpora and statistical analyses. The corpus frequency studied in the exploratory analysis proved to be another significant factor in adjectivization, as it functions both as a sign of adjectivization and as one of its causes.

The use of participles with the adverb *očen′* did not prove to be an independent factor for adjectivization, as stated in the syntactic approach. The analysis showed that, similar to gradable adjectives, a participle or a finite verbal form with a gradable meaning could easily be intensified by *očen*. However, only a qualitative cross-comparison of the frequency of the *očen* construction with a finite verb and with a participle showed that a participle may be adjectivized: The construction *očen* used with participles was more frequent than was that used with finite verbal forms when participles were adjectivized. The role of the verb semantics also proved to affect the frequency of the *očen′* used with participles: Participles derived from the base verbs with the semantic classes of psychological domains and the change of state showed the highest ratio in the *očen* construction in comparison to the participles with other semantic classes. In addition, the statistical test showed a significant association between the semantic classes of the participles and their ratios in the *očen′* construction, although it was not as strong as the association between the semantic classes of finite verbs and their ratios. It is not yet clear whether *očen′* generally prefers clear-cut adjectives as opposed to verbal forms, as this was not discussed in this dissertation.

The empirical analysis revealed that the pervasiveness of participles based on the rank frequency of the base verbs accounted for the ambiguity of participles. While the analysis confirmed some of the intuitions of the morphosemantic approach, such as the strong effect of the past tense[2] and the passive voice, alone or interacting with each other (Kustova, 2012; Kalakuckaja, 1971; Kolochkova, 2011), it also revealed some new findings that were not mentioned previously in the studies on adjectivization. First, the pervasiveness of participial forms and the rank of their corresponding base verbs strongly predicted adjectivization. Moreover, the observations of the distributions showed that the frequency rank of base verbs was associated with the morphological types and ambiguity of participles: The lower their rank, the higher the ratio of participles to verbal finite/infinitival forms, and the higher the number of ambiguous participles. The low-ranking verbs also tended to have more ambiguous past passive participles, while the high-ranking verbs appeared to have more present active participles. Another finding not discussed in the approaches to adjectivization concerns the effect of transitivity and aspect. Transitive perfective forms of participles and the perfective aspect, jointly with the ipm frequency of the base verbs, strongly predicted ambiguity (that is, ambiguous readings). However, the causes of these significant effects are not yet clear, and may be related to the frequent use of the adjectivized transitive perfective forms of participles in the disambiguated subcorpus of the RNC.

The disambiguation models (with morphological, syntactic, and weighted rules) based on the CG framework were cross-checked, and were also compared to a statistical (simple neural network) baseline model. The GSD model was introduced as an external baseline, which is a promising alternative to provide acceptable results in disambiguating at least one type of ambiguity. The CG-based models were built on the basis of (a) the framework of adjectivization[3] and (b) the corpus frequencies of adjectival and participial lemmas from Sharoff et al.'s (2013) dictionary implemented as weights in the morphological analyzer for Russian. Thus, the rules in the CG models expressed the properties of the syntactic context in which adjectivized and unambiguous participles were used, as well as the morphological properties of participles. The syntactic CG rules define specific (or fine-grained) and general (or coarse-grained) contexts surrounding an ambiguous word-from; the morphological CG rules do not specify any context except for several morphological rules combined with syntactic rules that specify the general context. All the morphological rules and a few syntactic rules contain the condition for weight comparison. The specific context covered in the syntactic CG rules consists of various immediate and remote constituents (adjuncts, adverbial modifiers, verb complements, and so forth) described in the rules after the instances found in the test corpora (the RNC, OpenCorpora, and the Yandex-1M corpus). The general context corresponds directly to the factors of adjectivization defined in the theoretical framework,[4] and lacks the broadness and detailed precision of the specific context. All the CG rules follow a deliberate order, namely specific context rules, general context rules, weighted morphological rules together with several joint morphological and syntactic rules, and the final weighted "select the reading with the highest weight" MAX rule, to ensure the reduction of ambiguity in a controlled and consistent way. That is, the rules disambiguate ambiguous word

[2] The past tense predicts unambiguous readings and the passive voice unambiguous ones.

[3] See Chapter 3.

[4] See Chapter 3.

forms used in a more specific syntactic context first and then proceed with the word forms used in a general (simple) syntactic context and/or by considering their morphological properties. The remaining instances of ambiguous participles are handled by the MAX rule in a straightforward manner by simply selecting morphological readings of the ambiguous word forms with the highest weight values.

The CG models that involve syntactic context (that is, the CTX-based models such as CTX, CTX+MAX, and CTX+$\mathrm{W_{MC}}$+$\mathrm{W_M}$+MAX) resolved ambiguity with more than 80% accuracy. The overall performance of these models was superior to the baseline statistical model, although their ambiguity ratio was higher than that of the GSD model, which did not leave ambiguous tags. The weighted model with morphological properties only ($\mathrm{W_M}$+$\mathrm{W_{MC}}$+MAX) and the MAX model fell within the range of 70%–71% accuracy, and showed inferior results in comparison to the CTX and the baseline models that fell within the range of 78–86% accuracy. Finally, the GSD model outperformed the weighted MAX and $\mathrm{W_M}$+$\mathrm{W_{MC}}$+MAX models without syntactic rules with the weighted averaged metric scores[5] of 78%, and might be combined with additional methods (such as features in Conditional Random Fields, or weights as unigram probabilities) to achieve scores comparable to those of the CTX models.

The flexibility of the CG framework allows for implementing corpus frequencies, and morphological and syntactic rules in the way that fine- and coarse-grained syntactic contexts and internal morphosyntactic properties are captured. More specifically, the syntactic context proved to be the most successful basis for the CG models. The rules based on morphological factors improved the performance of the syntactic CTX models, but these rules alone produced results inferior to the CG models with syntactic rules and the baseline model. When used in a separate model, these rules provided results comparable to the results of the MAX model. Thus, the optimal setting for distinguishing participles from adjectives is to use the rules based on the syntactic context in combination with morphological rules and weights. The weights on their own achieved around 70% precision simply by selecting the morphological readings with the highest weight value. The performance of weights might have been overshadowed by the limited size of the gold standard, and the performance may be better when using a larger set of evaluation corpora. In addition, differentiating among the weights of morphological types of participles (such as present active, past passive, and the like) might have led to a more accurate comparison of the weights for adjectival and participial readings.

3 Future directions

The dissertation provides a number of possibilities for future research with regard to the results discussed in the theoretical domain, the exploratory analysis, and the development of the disambiguation model.

[5] These metrics were weighted averaged precision, recall, f1-score, and accuracy.

As a type of conversion that is a universal process, adjectivization occurs in a number of unrelated language families, both synthetic and analytical, such as German, French, Romanian, Russian, Bulgarian, and English, among others (e.g., Valera, 2014; Štekauer et al., 2012; Don, 2003; Müller et al., 2015; Pšeničnaja, 2012). Possible further research could include expanding the scope from adjectivization to conversion, and exploring the common mechanisms that underlie conversion cross-linguistically. For example, one could analyze conversion in Russian in comparison to Bulgarian as a Slavic language with an analytical structure (*cf.* Manova, 2005). Conversion could also be studied cross-linguistically, and could involve both analytical European languages and creole languages (such as Jamaican and Gullah) that have limited or no derivational/inflectional morphology.

It would be equally interesting to focus on the lexical semantics of adjectivized participles and their relationship to the nouns they modify and regular adjectives. The productivity of certain types of adjectivized participles, presumably arising from the language user's need to convey an additional meaning not expressed by existing adjectives, has not been explored extensively in the existing literature. Finding linguistic and/or extra-linguistic reasons for the extension of meaning in adjectivized participles may clarify this issue.

Another interesting topic, although only related indirectly to adjectivization, is the compatibility of qualitative and relational adjectives (*cf.* Vol′f, 2002) with the abstract or concrete meaning of the head noun.[6] Developing a method that would allow for a distinction between relational and qualitative adjectives based on the meaning of the head noun could also reveal more about adjectivized participles. Modeling semantics for analyzing the polysemy of adjectives could be accomplished using Latent Semantic Analysis (LSA; Deerwester et al. 1990), the Correlated Occurrence Analogue to Lexical Semantics, or COALS (Rohde et al., 2006), or other, similar, vector-based methods.

Identifying directionality in the change in the semantics and grammatical meaning of an adjectivized participle in diachrony may establish the cause-effect relationship in the stages of the adjectivization that I defined in the synchronic analysis. As an example,[7] del Prado Martín and Brendel (2016) investigated whether the pattern of change in one Islandic case triggered the pattern of change in another (Granger-cause) using distance measures.

With regard to a practical application, implementing semantic information in the morphological transducer for annotation may advance CG-based disambiguation. The semantic information could be used as vector representation output from the neural network models, or taken from the lexicography databases for syntactic categories; for example, semantic features from the *Russian FrameBank* (Lyashevskaya, 2012; Lyashevskaya and Kashkin, 2015).

Using weights on a larger scale corpus and comparing their performance to a purely statistical model (such as the neural probabilistic model for morphology and POSs discussed by Cotterell and Schütze 2015) could provide a clearer picture of how the weights function, as well as their

[6] Mitrofanova (personal communication).

[7] Kapatsinski (personal communication).

advantages and disadvantages. Furthermore, separating the weights for each morphological type in a syntactic category may enable a more accurate comparison of weighted morphological analyses in disambiguation.

Bibliography

Afanas′ev, R. N. and Kobzareva, T. J. (2003). Intellektual′naja sistema predsintaksičeskogo analiza russkogo teksta (ISPA) [Intellectual system of pre-syntactic analysis of Russian text]. *Komp′juternaja lingvistika i intellektual′nye texnologii*, (pp. 5–10).

Ahmanova, A. S. (1984). *Slovar omonimov russkogo jazyka [Homonymy dictionary of Russian]*. Moscow: Soviet encyclopedia.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In J. Holub and J. Žďárek (Eds.), *Implementation and Application of Automata* (pp. 11–23). Berlin, Heidelberg: Springer Berlin Heidelberg.

Allen, J. F.and Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. R. (1995). The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 7(1), 7–48.

Allen, S. (2009). *Verb argument structure*, (pp. 217–236). Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Arakin, V. D. (2005). *Sravnitel′naja tipologija anglijskogo i russkogo jazykov [Comparative typology of English and Russian]*. Moscow: FIZMATLIT.

Babajceva, V. V. (1967). *Perexodnye konstrukcii v sintaksise [Transitive constructions in syntax]*. Voronež.

Bardina, T. K. (2003). *Problema leksiko-grammatičeskoj perexodnosti častej reči v sovremennom russkom jazyke [The problem of lexico-grammatical transitivity of parts of speech in modern Russian]*. PhD thesis, Volgogradskij gosudarstvennyj pedagogičeskij insitut, Volgograd.

Barnard, A. (2012). *Structure graph grammars and structure graph automata*. PhD thesis, University of Johannesburg.

Bauer, L. and Valera, S. (2005a). *Approaches to Conversion/Zero-derivation*. Waxmann: Münster.

Bauer, L. and Valera, S. (2005b). Conversion or zero-derivation: an introduction. *Approaches to conversion/zero-derivation*, (pp. 7–17).

Beard, R. (1998). Derivation. In *The Handbook of Morphology* (pp. 44–65). Blackwell Publishers Ltd.

Beesley, K. and Karttunen, L. (2003a). *Finite State Morphology*. CSLI Publications.

Beesley, K. and Karttunen, L. (2003b). *Two-Level Rule Compiler. Technical Report ISTL-92-2*. Technical report, Xerox Palo Alto Research Center, Palo Alto, California.

Belousov, V. N., Kovtunova, I., and Kručinina, I. (1989). *Kratkaja russkaja grammatika [A short grammar of Russian]*. Nauka, Moskva.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (Eds.), *Text, Speech and Dialogue* (pp. 247–256). Cham: Springer International Publishing.

Bergenholtz, H. and Agerbo, H. (2014). There is no need for the terms polysemy and homonymy in lexicography. *Lexikos*, 24, 27–35.

Berwick, R. C. and Weinberg, A. S. (1986). *The grammatical basis of linguistic performance: Language use and acquisition*. MIT press.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus Universitetsforlag.

Bick, E. (2009). Basic constraint grammar tutorial for cg-3 (vislcg3). *Southern Denmark University. CG-3 how-to*, 4, 2009.

Bick, E. and Didriksen, T. (2015). CG-3 – Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* (pp. 31–39).

Blox, M. J. (1983). *Teoretičeskaja grammatika anglijskogo jazyka: Učebnik [Theoretical grammar of English: a textbook]*. Vysšaja škola.

Bondarko, A. V. (1983). *Principy funkcional'noj grammatiki i voprosy aspektologii [Principles of functional grammar and problems of aspectology]*. Leningrad.

Bořkovec, V. Ž. (1976). Russian Verbs: The Question of Transitivity. *Russian Language Journal / Russkij jazyk*, 30(105), 1–7.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference* (pp. 224–231). Seattle, Washington, USA: Association for Computational Linguistics.

Briscoe, T., Grover, C., Boguraev, B., and Carroll, J. A. (1987). A Formalism and Environment for the Development of a Large Grammar of English. In *IJCAI*, volume 87 (pp. 703–708). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Brownlee, J. (2020a). How to calculate precision, recall, and f-measure for imbalanced classification. *URL: https://machinelearningmastery. com/precisionrecall-and-f-measure-for-imbalanced-classification*.

Brownlee, J. (2020b). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery.

Calude, A. S. and Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical transactions of the Royal Society of London*, 366(1567), 1101–1107.

Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i-Cancho, R., and Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech and Language*, 58, 19–50.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 1–8).: Association for Computational Linguistics.

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cotterell, R. and Schütze, H. (2015). Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1287–1292). Denver, Colorado: Association for Computational Linguistics.

Cousineau, D. (2020). How many decimals? Rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology*, 97, 102362.

Crystal, D. (2008). *A dictionary of linguistics and phonetics (The Language Library)*. Blackwell Publishing, John Wiley and Sons Incorporated.

Cumming, G., Fidler, F., Kalinowski, P., and Lai, J. (2012). The statistical recommendations of the American Psychological Association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, 64(3), 138–146.

Dalrymple, M. (2001). *Lexical functional grammar*. Brill.

de Gispert, A., Iglesias, G., Blackwood, G., Banga, E. R., and Byrne, W. (2010). Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-n Grammars. *Computational Linguistics*, 36(3), 505–533.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4585–4592). Reykjavik, Iceland: European Language Resources Association (ELRA).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391–407.

Dehaene, S. and Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1), 1–29.

del Prado Martín, F. M. and Brendel, C. (2016). Case and cause in Icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2421–2430).

DePoy, E. and Gitlin, L. N. (2016). Chapter 20 – Statistical Analysis for Experimental-Type Designs. In E. DePoy and L. N. Gitlin (Eds.), *Introduction to Research (Fifth Edition)* (pp. 282–310). St. Louis: Mosby, fifth edition edition.

Derbyshire, W. W. (1967). Verbal homonymy in the Russian Language. *Canadian Slavonic Papers*, 9(1), 131–139.

Didriksen, T. and ApS, G. (2007). *Constraint Grammar Manual*.

Dokulil, M. (1968). Zur Frage der Konversion und verwandter Wortbildungsvorgänge und-beziehungen. *Travaux linguistiques de Prague*, 3, 215–239.

Don, Z. M. (2003). Language-dialect code-switching: Kelantanese in a multilingual context. *Multilingua*, 22(1), 21–40.

Dressler, W. U. (2005). *Word-Formation in Natural Morphology*, (pp. 267–284). Springer Netherlands: Dordrecht.

Dressler, W. U. and Manova, S. (2002). Conversion vs. modification and subtraction. Paper Presented at the Seminar on Conversion/Zero-Derivation.

Droste, M., Kuich, W., and Vogler, H. (2009). *Handbook of weighted automata*. Number 1. Springer Science and Business Media.

Dunlap, J. (1981). Russian verbal aspect and transitivity. Master's thesis, University of Alberta.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64.

Ferrer-i-Cancho, R. (2016). The meaning-frequency law in Zipfian optimization models of communication. *Glottometrics*, 35, 28–37.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications Ltd., 4th edition.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.

Forsyth, J. (1970). *A grammar of Aspect: Usage and Meaning in the Russian verb*. Cambridge University Press.

Gak, V. G. (2002). *Lingvističeskij ènciklopedičeskij slovar′ [Linguistic encyclopedic dictionary]*, chapter Transpozicija [Transposition]. Bol′šaja Rossijskaja ènciklopedija: Moscow.

Gazdar, G., Klein, E., and Pullum, G. (1985). *Generalized Phrase Structure Grammar*. Massachusetts: Blackwell and Cambridge.

Givón, T. (1991). Markedness in Grammar: Distributional, Communicative and Cognitive Correlates of Syntactic Structure. *Studies in Language*, 15.

Greenbaum, S. (1996). *The Oxford English Grammar*. Oxford University Press.

Hanneforth, T. (2008). *Finite-State Methods and Natural Language Processing: 6th International Workshop, FSMNLP 2007. Revised Papers*. Universitätsverlag Potsdam.

Hansen, B., Hansen, K., Neubert, A., and Schentke, M. (1982). *Englische Lexikologie*. Enzyklopädie, Leipzig edition.

Haynes, W. (2013). *Wilcoxon Rank Sum Test*, (pp. 2354–2355). Springer New York: New York.

Joshi, A. K. (1985). *Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?*, (pp. 206–250). Studies in Natural Language Processing. Cambridge University Press.

Kalakuckaja, L. P. (1971). *Adjektivatsija prichastij v sovremennom russkom literaturnom jazyke [Adjectivization in modern Russian literary language]*. Moscow.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING ’90 (pp. 168–173). USA: Association for Computational Linguistics.

Karlsson, F. (1992). SWETWOL: A Comprehensive Morphological Analyser for Swedish. *Nordic Journal of Linguistics*, 15(1), 1–45.

Karlsson, F. (1995). *Designing a parser for unrestricted text*, (pp. 1–40). De Gruyter Mouton: Berlin, Boston.

Karttunen, L. (1993). *Finite-State Lexicon Compiler*. Technical report, Xerox Palo Alto Research Center, Palo Alto, California.

Karttunen, L. (1994). Constructing Lexical Transducers. In *The Proceedings of the 15th International Conference on Computational Linguistics COLING 94*, volume 1 (pp. 406–411).

Karttunen, L., Koskenniemi, K., Kaplan, R., et al. (1987). A compiler for two-level phonological rules. *Tools for morphological analysis*.

Katamba, F. (1993). Productivity in Word-Formation. In *Morphology* (pp. 65–85). Springer.

Keleg, A., Tyers, F., Howell, N., and Pirinen, T. A. (2020). An unsupervised method for weighting finite-state morphological analyzers. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 3842–3850).

Klyšinskij, E. and Rysakov, S. V. (2015). Statističeskie metody snjatija omonimii [Statistical methods of homonymy resolution]. *Novye informacionnye texnologii v avtomatizirovannyx sistemax*, (pp. 555–563).

Knight, K. and May, J. (2009). Applications of weighted automata in natural language processing. In *Handbook of Weighted Automata* (pp. 571–596). Springer.

Kolochkova, O. V. (2011). *Ad″ektivirovannye realizacii pričastij v sovremennom russkom jazyke [Adjectivized realizations of participles in modern Russian]*. PhD thesis, Saint Petersburg State University.

Koskela, A. A. and Murphy, M. L. (2006). Polysemy and homonymy. In K. Brown, A. H. Anderson, L. Bauer, M. Berns, G. Hirst, and J. Miller (Eds.), *Encyclopedia of language and linguistics (2nd ed.)*, volume 9 (pp. 742–744). Elsevier.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*, volume 11. University of Helsinki, Department of General Linguistics Helsinki, Finland.

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 4th Nordic Conference of Computational Linguistics (NODALIDA 1983)* (pp. 145–154).

Koskenniemi, K. (1990). Finite-state parsing and disambiguation. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90 (pp. 229–232). USA: Association for Computational Linguistics.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583–621.

Kubrjakova, E. S. (1974). Derivatsija, transpozitsija, konversija [Derivation, transposition, conversion]. *Voprosy jazykoznanija*, 5, 64–76.

Kumar, S. and Byrne, W. (2003). A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 142–149).

Kumar, S., Deng, Y., and Byrne, W. (2006). A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering*, 12(1), 35–75.

Kustova, G., I. (2001). Database of verbs and their semantic roles. This is a database of verbal lemmas and their semantic roles of 05.12.2001, granted me by Galina Kustova.

Kustova, G., I. (2012). Semantic types and semantic functions of the adjectivized participles. *Komp'juternaja lingvistika i intellektual'nye texnologii. Po materialam meždunarodnoj konferencii "Dialog 2012"*, 1(11(18)), 352–361.

LaVange, L. M. and Koch, G. G. (2006). Rank Score Tests. *Circulation*, 114(23), 2528–2533.

Letučij, A. B. (2018). Predikativy [Predicates]. *Materialy dlja proekta korpusnogo opisanija russkoj grammatiki*, 3.

Levine, R. and Meurers, D. (2006). Head-Driven Phrase Structure Grammar Linguistic Approach, Formal Foundations, and Computational Realization. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*. Oxford: Elsevier.

Lieber, R. (2005). English Word-Formation Processes. In *Handbook of Word-Formation*. Dordrecht: Springer Netherlands.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., and Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2), 374–397.

Lindberg, N. and Eineborg, M. (1998). Learning constraint grammar-style disambiguation rules using inductive logic programming. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2* (pp. 775–779).

Linden, K. and Pirinen, T. (2009). Weighted finite-state morphological analysis of Finnish compounding with HFST-LEXC. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, (pp. 89–95).

Lindén, K., Pirinen, T., et al. (2009a). Weighting Finite-State Morphological Analyzers using HFST tools. In *FSMNLP*, volume 13: Citeseer.

Lindén, K., Silfverberg, M., and Pirinen, T. (2009b). HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. In C. Mahlow and M. Piotrowski (Eds.), *State of the Art in Computational Morphology* (pp. 28–47). Berlin, Heidelberg: Springer Berlin Heidelberg.

Listenmaa, I. (2019). *Formal Methods for Testing Grammars*. PhD thesis, Chalmers University of Technology and University of Gothenburg.

Loftsson, H. (2007). *Tagging and parsing Icelandic text*. PhD thesis, University of Sheffield, UK.

Lopatin, V. V. (1966). Adjektivatsija prichastij v ee otnoshenii k slovoobrazovaniju [Adjectivization of participles and word-formation]. *Voprosy jazykoznanija [Problems of Linguistics]*, (5), 37–48.

Lundquist, B., Iordachioaia, G., Roy, I., and Takamine, K. (2013). The category of participles. *Gianina Iordachioaia, Isabelle Roy and Kaori Takamine (eds.), Categorization and category change*, (pp. 11–32).

Lyashevskaya, O. (2012). Dictionary of valencies meets corpus annotation: a case of Russian framebank. In *Proceedings of the 15th EURALEX International Congress*, volume 15 Oslo, Norway: Oslo University.

Lyashevskaya, O. and Kashkin, E. (2014). Evaluation of frame-semantic role labeling in a case-marking language. *Computational linguistics and intellectual technologies*, 13(20), 362–378.

Lyashevskaya, O. and Kashkin, E. (2015). Framebank: a database of Russian lexical constructions. *Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015*, 542, 337–348.

Malmkjær, K. (2010). *The Routledge linguistics encyclopedia*. London: Routledge, 3 edition.

Manova, S. (2005). Towards a Theory of Conversion in Slavic: Evidence from Bulgarian, Russian and Serbo-Croatian. *Glossos*, 6, 1–27.

Manova, S. (2011). *Understanding Morphological Rules: With Special Emphasis on Conversion and Subtraction in Bulgarian, Russian and Serbo-Croatian*. Studies in Morphology. Springer Netherlands.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word Formation*. München: Beck.

Marcus, M. P. (1980). *Theory of Syntactic Recognition for Natural Languages*. Cambridge, MA, USA: MIT Press.

Marek, T. (2006). Analysis of German compounds using weighted finite-state transducers. *Bachelor thesis, University of Tübingen*.

Matthews, P. H. (2014). *The Concise Oxford Dictionary of Linguistics*. OUP Oxford.

Mohri, M., Pereira, F., and Riley, M. (2008). *Speech Recognition with Weighted Finite-State Transducers*, (pp. 559–584). Springer Berlin Heidelberg: Berlin, Heidelberg.

Müller, N., Gil, L. A., Eichler, N., Geveler, J., Hager, M., Jansen, V., Patuto, M., Repetto, V., and Schmeißer, A. (2015). *Code-Switching: Spanisch, Italienisch, Französisch. Eine Einführung*. Narr Francke Attempto Verlag.

Müller, S. (2013). *Head-Driven Phrase Structure Grammar: Eine Einführung*. Number 17 in Stauffenburg Einführungen. Tübingen: Stauffenburg Verlag, 3 edition.

Nikitevič, V. M. (1985). *Osnovy nominativnoj derivacii [Foundations of nominative derivation]*. Vyšèjšaja škola.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Panman, O. (1982). Homonymy and polysemy. *Lingua*, 58(1-2), 105–136.

Panova, G. I. (2010). *Morfologija russkogo jazyka: enciklopedičeskij slovar′-spravočnik [Morphology of Russian: encyclopedic dictionary]*. KomKniga.

Parmenova, T. V. (2002). Functional approach to the study of grammar at school (about one of the ways to modernization). *Russian*, (24).

Peirsman, Y. and Geeraerts, D. (2006). Metonymy as a prototypical category. *Cognitive Linguistics*, 17(3), 269–316.

Pereira, F., Riley, M., and Sproat, R. (1994). Weighted rational transductions and their application to human language processing. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11* New Jersey.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). Istanbul, Turkey: European Language Resources Association (ELRA).

Petruhina, V. E. (2006). Aktual′nye voprosy sistemnogo slovoobrazovanija [Recent issues of systemic word-formation]. *Slavistika: sinxronija i diaxronija. Sbornik statetj k 70-letiju I. S. Uluxanova.*, (pp. 142–154).

Piantadosi, S. (2012). Approximate number from first principles. Manuscript under review.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review*, 21(5), 1112–1130.

Pin, J.-É. (2016). *A tutorial on sequential functions*. LIAFA, CNRS and University Paris 7, Amsterdam.

Plungjan, V. A. (2010). Pričastija i psevdopričastija v russkom jazyke: o granicax variativnosti [Participles and presudoparticiples in Russian: boundaries of variation]. Report presented on 26.02.2010 in Oslo.

Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Pulkina, I. M. (1987). *A short Russian reference grammar*. Progress.

Pšeničnaja, A. (2012). Obsčaja xarakteristika konversii kak sposoba slovoobrazovanija (na materiale russkogo i francuzskogo jazykov) [General properties of conversion as means of word-formation (based on the material of Russian and French)]. *Izvestija Južnogo federal′nogo universiteta*, (1), 145–150.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive grammar of the English language*. London: Longman.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranganathan, P., Pramesh, C. S., and Aggarwal, R. (2017). Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research*, 8(3), 148.

Rassudova, O. P. (1982). *Upotreblenie vidov glagola [Use of verbal aspects]*. Moscow: Russkij yazyk.

Reynolds, R. (2016). *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. PhD thesis, UiT The Arctic University of Norway.

Rohde, D. L., Gonnerman, L. M., and Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8(627-633), 116.

Sag, I. A. (2003). Coordination and underspecification. In *Proceedings of the 9th HPSG conference* (pp. 267–291).

Sak, H., Saraclar, M., and Gungor, T. (2012). Morpholexical and discriminative language models for Turkish automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8), 2341–2351.

Say, S. (2016). Pričastie (v pečati) [Participles (to appear)]. *Materialy k korpusnoj grammatike russkogo jazyka*.

Schiller, A. (2005). German compound analysis with wfst. In *International Workshop on Finite-State Methods and Natural Language Processing* (pp. 239–246).: Springer.

Schlücker, B. (2019). *Complex Lexical Units. Compounds and Multi-Word Expressions*. De Gruyter.

Schönefeld, D. (2005). Zero-derivation–functional change–metonymy. *Approaches to conversion/zero-derivation*, (pp. 131–159).

Sharoff, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Creating General-Purpose Corpora Using Automated Search Engine Queries. In *Wacky! Working papers on the Web as Corpus* (pp. 63–98). Bologna: GEDIT.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and Evaluating a Russian Tagset. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*.

Sharoff, S. and Nivre, J. (2011). The proper place of men and machines in language technology. *Processing Russian without any Linguistic Knowledge. Computational Linguistics and Intelligent Technologies*.

Sharoff, S., Umanskaya, E., and Wilson, J. (2013). *A Frequency Dictionary of Russian: Core Vocabulary for Learners*.

Sičinava, D. V. (2018). Narečija [Adverbs]. *Materialy k korpusnoj grammatike russkogo jazyka. Vypusk III: Časti reči i leksiko-grammatičeskie klassy*, (pp. 108–135).

Sjöbergh, J. and Kann, V. (2004). Finding the correct interpretation of Swedish compounds, a statistical approach. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* Lisbon, Portugal: European Language Resources Association (ELRA).

Smirnickij, A. I. (1954). Po povodu konversii v anglijskom jazyke [Conversion in English]. *Inostrannye jazyki v škole*, (3).

Smit, P., Virpioja, S., Kurimo, M., et al. (2017). Improved Subword Modeling for WFST-Based Speech Recognition. In *Proceedings of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017) : Situated Interaction* (pp. 2551–2555).

Spoustová, D. J., Hajič, J., Raab, J., and Spousta, M. (2009). Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 763–771). Athens, Greece: Association for Computational Linguistics.

Stein, G. (1977). The place of word-formation in linguistic description. *Perspektiven der Wortbildungsforschung. Beiträge zum Wuppertaler Wortbildungskolloquium vom 9.–10. Juli 1976. Anläßlich des 70. Geburtstages von Hans Marchand am 1. Oktober 1977*, (pp. 219–235).

Straka, M., Hajic, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290–4297).

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Baltimore, Maryland: Association for Computational Linguistics.

Tanenhaus, M. K. and Sedivy, J. C. (2001). *MIT Encyclopedia of Cognitive Science*, chapter Ambiguity. MIT Press: Cambridge, Mass.

Tapanainen, P. (1996). *The constraint grammar parser CG-2*. Department of General Linguistics, University of Helsinki.

Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge University Press.

Tsarfaty, R. (2013). A Unified Morpho-Syntactic Scheme of Stanford Dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 578–584). Sofia, Bulgaria: Association for Computational Linguistics.

Tyers, F. and Reynolds, R. (2015). A preliminary constraint grammar for Russian. *Proceedings of the Workshop on "Constraint Grammar – methods, tools, applications" at NODALIDA 2015*.

Valera, S. (2014). Conversion. In *The Oxford Handbook of Derivational Morphology* (pp. 154–168). Oxford University Press.

Valera, S. (2015). *Conversion*, (pp. 322–339). De Gruyter Mouton: Berlin, Boston.

van Halteren, H., Zavrel, J., and Daelemans, W. (2001). Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems. *Computational linguistics*, 27(2), 199–230.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, 67(6), 1176–1190.

van Rooij, M., Nash, B., Rajaraman, S., and Holden, J. (2013). A fractal approach to dynamic inference and distribution analysis. *Frontiers in Physiology*, 4, 1.

Černega, L. V. (2009). *Pričastija i ix adjektivirovannye realizacii [Participles and their adjectivized realizations]*. PhD thesis, Taganrogskij gosudarstvennyj pedagogičeskij institut.

Vinogradov, V. V. (1954). Nekotorye zadači izučenija sintaksisa prostogo predloženija [Several tasks of studying syntax of a simple sentence]. *Grammatika russkogo jazyka. Sintaksis*, 2(1).

Vinogradov, V. V. (1972). *Russkij jazyk (Grammatičeskoe učenie o slove) [Russian language (Grammatical study of a word)]*. Moskva.

Vinogradov, V. V. (1975). *Izbrannye trudy. Issledovanija po russkoj grammatike [Selected works. Research studies on Russian grammar]*. Moscow: Nauka.

Vinogradov, V. V., Istrina, E. S., and Barkhudarov, S. G. (1960). *Grammatika russkogo jazyka [Russian grammar]*, volume 1. Akademija Nauk SSSR.

Vogel, P. M. (1996). *Wortarten und Wortartenwechsel. Zu Konversion und verwandten Erscheinungen im Deutschen und in anderen Sprachen*, volume 39. Berlin and New York: Walter de Gruyter.

Volʹf, E. (2002). *Funkcionalʹnaja semantika ocenki [Functional semantics of judgement]*. Editorial URSS.

Voutilainen, A. (1994). Designing a parsing grammar. In *Finite-State Language Processing*.

Voutilainen, A. and Tapanainen, P. (1993). Ambiguity resolution in a reductionistic parser. *EACL*.

Šaxmatov, A. A. and Istrina, E. S. (1963). *Sintaksis russkogo jazyka [Syntax of the Russian language]*, volume 41. Mouton.

Štekauer, P., Valera, S., and Kőrtvélyessy, L. (2012). *Word-Formation in the World's Languages: A Typological Survey*, chapter Word-formation without addition of derivational material and subtractive word-formation, (pp. 213–236). Cambridge University Press: Cambridge.

Švedova, N. (1980). *Russkaja grammatika [Russian grammar]*, volume 1. Nauka.

Žerebilo, T. V. (2010). *Slovar′ lingvističeskix terminov [Dictionary of linguistic terms].* Piligrim.

Wiechetek, L. (2018). *When grammar can't be trusted - Valency and semantic categories in North Sámi syntactic analysis and error detection.* PhD thesis, University of Tromsø.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

Wilson, R. A. and Keil, F. C. (2001). *The MIT encyclopedia of the cognitive sciences.* MIT press.

Wmffre, I. (2013). *Dynamic Linguistics: Labov, Martinet, Jakobson and Other Precursors of the Dynamic Approach to Language Description.* Peter Lang.

Yli-Jyrä, A. (2014). Weighted automata. TU Darmstad.

Zajic, D., Dorr, B., and Schwartz, R. (2002). Automatic Headline Generation for Newspaper Stories. In *Proceedings of the ACL-02 Workshop on Text Summarization (DUC 2002)* (pp. 78–85). Philadelphia, PA: Association for Computational Linguistics.

Zaliznjak, A. A. (2003). *Grammatičeskij slovar′ russkogo jazyka [Grammatical Dictionary of the Russian Language].* Russkie slovari.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *LREC*, volume 2008 (pp. 28–30).

Zemskaja, E. A. (2008). *Sovremennyj russkij jazyk. Slovoobrazovanie: Ušebnoe posobie [Modern Russian. Word-formation: a textbook].* Moscow: Flinta, Nauka.

Zipf, G. K. (1945). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, 33(2), 251–256.

Zipf, G. K. (1999). *The psycho-biology of language: An introduction to dynamic philology*, volume 21. Psychology Press.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books.

Appendix A

Methodology

1 Basic definitions for weighted transducers

Semiring	**Interpretation**	**Weight of transition**	$\bigoplus$	$\bigotimes$
Probability	probability	$[0, 1]$	$+$	$\cdot$
Log	log-probaility	$\mathbb{R} \cup +\infty$	$\oplus log$	$+$
Tropical	best path log-probability (minimum weight value)	$\mathbb{R}^+ \cup \infty$	min	$+$
Polar	best path log-probability (maximum weight value)	$R \cup \infty$	max	$+$
Viterbi	best path probability	$[0, 1]$	max	$\cdot$
Max-min	best path probability	$[0, 1]$	max	min
Boolean	if string is in language	$0, 1$	$\bigvee$	$\bigwedge$
String	a string	$A*$ string	longest common prefix $\bigwedge$	string concatenation

Table A.1: Types of semirings and interpretations of the weight of the global path, as well as its binary addition $\oplus$ and multiplication $\otimes$ operations, which can be used as other types of operations depending on the type of semiring and the interpretation of the global path weight. The weight of transition represents the value of the weight selected by the algorithm using a given semiring.

A tropical semiring is used in algorithms to select the best path with the lowest weight when there are several paths or transitions that differ only in terms of weight. The calculation to obtain the value of the weight of the global path is performed using the operations *min* and $+$.

2 Files and commands

The generator can be found in the file `generator-raw-gt-desc.hfst` and is run via the alias command `hdrusStress`. The command `hfst-lookup` in `hfst-lookup itest.hfst` looks up and outputs morphological readings.The weighted finite-state transducer is located in the file `analyser-disamb-gt-desc.hfstol`. One compiles these tools following the instructions available at https://github.com/giellalt/lang-rus. Alternatively, one may refer to the ready-complied analyzers for text analysis that can be downloaded at https://giellalt.uit.no/ling/LinguisticAnalysis.html. The finite-state and CG-based analyzers and other tools, as well as language resources for Russian, are available at https://github.com/giellalt/lang-rus. The installation instructions are available at https://github.com/giellalt/lang-rus/blob/develop/INSTALL.

Appendix B

Ambiguity

Table B.1 presents the synthesis of Tyers and Reynolds's (2015) classification of ambiguity, as well as the descriptions and examples listed by Ahmanova (1984); Vinogradov (1954); Derbyshire (1967); Vinogradov et al. (1960). It provides examples of POS ambiguity and two subtypes of paradigmatic ambiguity, namely partial and full ambiguity.

Part of speech ambiguity	Partial paradigmatic ambiguity	Full paradigmatic ambiguity
	Homonyms	
1. Unrelated morphological forms and meanings • Verbs with adverbs, e.g., *počti* from *počtit′* ‘honor-N’ and *počti* ‘almost-ADV’ • Imperative verbs with numerals, e.g.,*pjat′* from *pjatit′* ‘move backward-V’ and ‘five-NUM’; *tri* from *teret′* ‘wipe-V-IMP’ and *tri* ‘three-NUM’ • Imperative verbs with pronouns, e.g., *moj* from *myt′* ‘wash-V’ and ‘my-PRO’ 2. Related morphological forms and meanings • Adjectives with nouns, e.g., ‘ice-ADJ’ versus ‘ice-cream-N’ • Participles with adjectives, e.g., *blestjaščij* ‘shining-PTCP’ versus ‘brilliant-ADJ’ • Past tense forms with the short forms of adjectives:, e.g., *smel* from *smet′* ‘dare-V’ and a short form *smelyj* ‘daring-ADJ’	• Verbs agreeing in the present/future conjugation but not in the infinitive form, e.g., *dobreju* from *dobrit′* ‘finish shaving’ and *dobret′* ‘become kinder’ • Verbs agreeing in the first person singular, e.g., *leču* from *lečit′* ‘treat’ and *letet′* ‘fly’; *melju* from *melit′* (‘rub with chalk’ or ‘make small’) versus *molot′* ‘mill’ • A clash of imperatives: *vej* from *vejat′* ‘fan’ versus *vit′* ‘twist’ • Verbs agreeing in the infinitive form but not in present/future conjugations, e.g., *žat′*: ‘reap’ versus ‘squeeze’, *klepat′*: ‘rivet’ versus ‘slander’, and *zret′*: ‘ripen’ versus ‘behold’	• Verbs agreeing with verbs: vowel/consonant alternations, no aspect match: suffix -*yvat′*–*ivat′* + consonant alternation(s > š), or vowel alternation (o > a) in a verb stem, e.g., *domešivat′* (imperfective) from *domešat′* ‘finish mixing’ and *domesit′* ‘finish kneading’ (perfective), *zasalivat′* from *zasalit′* ‘soil’ versus *zasolit′* ‘salt down’ (imperfective), and *zakapyvat′* from *zakopat′* ‘dig’ versus *zakapat′* ‘begin to drip’ • Homonymous prefixes: no aspect match, e.g., *sxodit′* ‘go to’ (perfective) and ‘go down’ (imperfective) • Aspect match: morphemes (that is, prefixes such as *c*-, *po*-, *ob*-, *na*-, and *pere*-) can develop homonymy (Vinogradov, 1972), e.g., *na*-: *nakolot′* with quantitative and spacial meanings (‘chop wood/carve a pattern’ and ‘pin up a badge’), *nastroit′* (‘tune up chords’ and ‘build (houses)’), *pro*-: *prosmotret′* (‘to watch until the end’, ‘look through’ and ‘overlook’), and *za*-: *zažit′* (‘heal a wound’ and ‘begin to live’)
	Homographs	
• Past tense forms agreeing with feminine/masculine/neuter nouns, e.g., *nachálo* (‘beginning’) and *náchalo* (‘it started’), and *žíla* (‘vein’) and *žilá* (‘[she] lived’) • Verbs in the first person singular with masculine nouns in the dative singular or feminine nouns in the accusative singular, e.g., *béregu* (‘bank’) and *beregú* (‘I keep’), and *prístan′* (‘harbor’) and *pristán′* (‘Stick to’) • Feminine accusative adjectives in the singular and the first person singular form of verbs, e.g., *celúju* (‘I kiss’) and *céluju* (‘the whole’)	• Verbs agreeing in several forms of the present tense but not in the infinitive form, e.g., *krojú* from *krojit′* (‘cut’) and *krója* from *kryt′* (‘cover’)	• Nouns agreeing with nouns, e.g., *zamók* (‘lock’) and *zámok* (‘castle’), and *muká* (‘flour’) and *múka* (‘suffering’) • Verbs coinciding graphically in infinitive forms and in all conjugations: aspect match, as in *zapáxnut′* (‘smell’) and *zapaxnút′* (‘wrap’) (perfective), or no aspect match as in *srezát′* (‘cut off’, imperfective) and *srézat′* (‘cut’, perfective), *spešít′* (‘hurry’, imperfective) and *spéšit′* (‘dismount’, perfective), *zasypát′* (‘fall asleep’, imperfective) and *zasýpat′* (‘fill in’, perfective), and *napadát′* (‘attack’, imperfective) and *napádat′* (‘fall down’, perfective)
	Polysemes	
	• *boltat′* (‘stir’ and ‘chatter’), *ostrit′* (‘sharpen’ and ‘crack jokes’)	

Table B.1: Classification of morphosyntactic ambiguity.

Appendix C

Conversion

Table C.1 briefly describes the types of conversion defined by Manova (2011).

Linguistic level	**Type of conversion**	**Properties**	**Examples**
Derivation	Word-class changing conversion	Use of addition, substitution and deletion of inflection Word- (the most prototypical), stem- or root-based (the least prototypical).	*slabyj* ‘weak-ADJ’ ⇒ *slabit′* ‘weaken-V’, English *to walk*-V ⇒ *a walk-N*
	Word-class preserving conversion	No derivational affixes, addition or deletion of inflections, non-prototypical, word-, stem- or root-based	*garden*-N ⇒ *garden-er*-N (Manova, 2011: 97)
Inflection	Formal conversion	Non-prototypical inflection. Word forms expressing suffixless gender and aspect through a paradigmatic change.	*bog* ‘God-N.M’ ⇒ *bog-in-ja* ‘Goddess-N.F’ (Manova, 2011: 108)
	Syncretism	Different syntactic function and agreement. Identical word-class and paradigm. No paradigmatic or semantic change.	*stol* ‘table.NOM.SG ’ = ‘table.ACC.SG’; *stol-y* ‘table.NOM.PL’ = ‘table.ACC.PL’; *cen-e* ‘price.DAT.SG ’ = ‘price.LOC.SG’; *cen-y* ‘price.GEN.SG’ = ‘price.ACC.PL’ (Manova, 2011: 59)
Syntax	Syntactic conversion	No affixation. Change of syntactic functions. Different word class.	*sladk-oe* ‘sweet-ADJ.NEUT’ ⇒ *sladk-oe* ‘dessert-N.NEUT’ (Manova, 2011: 112)

Table C.1: Manova’s (2011) classification of conversion lists the linguistic level, the types of conversion, and properties of these types illustrated by examples.

Appendix D

Adverbial *očen′* construction

1 Distributions

Table 1 presents the distributions of the raw frequencies of the present participles (column **Fpresp**) and present indicative verbs (column **Fv**). It also shows the ratio of the construction *očen′* + PRESP (column ***očen′* RatioPresp**) and *očen′* + PRS V (column ***očen′* RatioV**) for each verbal lemma. The column **Idx** shows the increasing values representing the differences between the ratios of *očen′* + PRESP and *očen′* + PRS V constructions for each verbal lemma.

#	Lemma	FreqPresp	***očen′*** **FreqPresp**	FreqV	***očen′*** **FreqV**	***očen′*** **RatioPresp**	***očen′*** **RatioV**	Idx
1	*nravit′sja*	44	0	83093	10820	0.00000	**0.13022**	-0.130215542
2	*volnovat′sja*	50	0	2930	267	0.00000	**0.09113**	-0.09112628
3	*ljubit′*	15316	237	164104	16210	0.01547	**0.09878**	-0.083304809
4	*radovat′*	803	1	22282	1718	0.00125	**0.07710**	-0.075857264
5	*uvažat′*	3158	1	7497	544	0.00032	**0.07256**	-0.072245702
6	*smaxivat′*	45	0	697	50	0.00000	**0.07174**	-0.071736011
7	*pereživat′*	692	2	14745	851	0.00289	**0.05771**	-0.054824306
8	*cenit′*	382	1	21390	1079	0.00262	**0.05044**	-0.047826332
9	*bodrit′*	307	1	895	35	0.00326	**0.03911**	-0.035848816
10	*obodrjat′*	176	1	147	6	0.00568	**0.04082**	-0.035134508
11	*vdoxnovljat′*	494	14	2942	180	0.02834	**0.06118**	-0.032842788
12	*intrigovat′*	492	10	366	18	0.02033	**0.04918**	-0.028855125
13	*motivirovat′*	112	2	1608	71	0.01786	**0.04415**	-0.026297086
14	*volnovat′*	2023	13	11817	385	0.00643	**0.03258**	-0.026154081
15	*osvežat′*	452	2	1720	47	0.00442	0.02733	-0.022900803
16	*zatjagivat′*	28	0	2723	53	0.00000	**0.01946**	-0.019463827
17	*bojat′sja*	726	3	49930	1160	0.00413	**0.02323**	-0.019100294
18	*interesovat′*	28198	8	34429	653	0.00028	**0.01897**	-0.018682861
19	*somnevat′sja*	1109	3	13635	290	0.00271	**0.02127**	-0.018563654
20	*napominat′*	12520	208	46190	1603	0.01661	**0.03470**	-0.018091063
21	*rekomendovat′*	5251	1	99236	1705	0.00019	**0.01718**	-0.016990825
22	*pomogat′*	8293	8	150185	2303	0.00096	**0.01533**	-0.014369752
23	*interesovat′sja*	4096	11	13256	207	0.00269	**0.01562**	-0.012930023

Table D.1 – continued on the next page

#	**Lemma**	**FreqPresp**	***očen′* FreqPresp**	**FreqV**	***očen′* FreqV**	***očen′* RatioPresp**	***očen′* RatioV**	**Idx**
24	*uspokaivat′*	3994	6	5128	65	0.00150	**0.01268**	-0.011173254
25	*pugat′*	2585	16	7836	128	0.00619	**0.01633**	-0.01014531
26	*obnadeživat′*	660	28	457	24	0.04242	**0.05252**	-0.010092169
27	*mešat′*	2986	5	30798	345	0.00167	**0.01120**	-0.009527545
28	*nuždat′sja*	10114	10	50708	530	0.00099	**0.01045**	-0.009463271
29	*šokirovat′*	423	0	572	5	0.00000	**0.00874**	-0.008741259
30	*privlekat′*	3263	1	37536	333	0.00031	**0.00887**	-0.008565017
31	*razbirat′sja*	378	1	12359	134	0.00265	**0.01084**	-0.008196799
32	*počitat′*	2368	39	2244	53	0.01647	**0.02362**	-0.007148944
33	*stradat′*	15831	1	33185	237	0.00006	**0.00714**	-0.007078614
34	*vpečatljat′*	1498	62	6514	314	0.04139	**0.04820**	-0.006815351
35	*različat′sja*	1766	3	13523	108	0.00170	**0.00799**	-0.006287639
36	*žaždat′*	1815	1	2864	19	0.00055	**0.00663**	-0.006083114
37	*podxodit′*	19977	79	133325	1314	0.00395	**0.00986**	-0.005901068
38	*stimulirovat′*	5947	1	16384	72	0.00017	**0.00439**	-0.004226379
39	*vlijat′*	13824	2	72772	264	0.00014	**0.00363**	-0.003483093
40	*otličat′sja*	14714	30	160795	687	0.00204	**0.00427**	-0.002233646
41	*razvivat′*	10410	2	14894	33	0.00019	**0.00222**	-0.002023534
42	*sposobstvovat′*	10141	5	85047	212	0.00049	**0.00249**	-0.001999691
43	*podderživat′*	14620	2	49794	97	0.00014	**0.00195**	-0.001811227
44	*zaviset′*	6184	3	212880	321	0.00049	**0.00151**	-0.001022769
45	*želat′*	23917	7	75145	91	0.00029	**0.00121**	-0.000918313
46	*umet′*	5296	1	52446	56	0.00019	**0.00107**	-0.000878943
47	*stremit′sja*	5990	1	53452	52	0.00017	**0.00097**	-0.000805891
48	*idti*	15381	5	235019	258	0.00033	**0.00110**	-0.000772707
49	*potrjasat′*	19693	6	1098	1	0.00030	**0.00091**	-0.00060607
50	*goret′*	10944	4	11341	11	0.00037	**0.00097**	-0.000604435
51	*sootvetstvovat′*	34333	5	79297	53	0.00015	**0.00067**	-0.000522741
52	*vladet′*	6619	0	19225	9	0.00000	**0.00047**	-0.00046814
53	*uznavat′*	26	0	2635	1	0.00000	**0.00038**	-0.000379507
54	*obitat′*	3160	0	9862	1	0.00000	**0.00010**	-0.000101399
55	*rabotat′*	74629	1	329941	25	0.00001	**0.00008**	-6.23715E-05
56	*pol′zovat′sja*	5075	1	98693	24	0.00020	**0.00024**	-4.6134E-05
57	*upravljat′*	8651	1	10873	1	**0.00012**	0.00009	2.36226E-05
58	*trebovat′*	59558	4	176264	5	**0.00007**	0.00003	3.87949E-05
59	*oxranjat′*	13822	1	2851	0	**0.00007**	0.00000	7.23484E-05
60	*proxodit′*	15990	2	133961	7	**0.00013**	0.00005	7.28242E-05
61	*stoit′*	8510	2	470393	27	**0.00024**	0.00006	0.000177619
62	*vyzyvat′*	13773	3	89787	2	**0.00022**	0.00002	0.000195543
63	*blestet′*	2030	8	2171	8	**0.00394**	0.00368	0.000255949
64	*vozbuždat′*	1962	26	2710	35	**0.01325**	0.01292	0.000336655
65	*ponimat′*	4405	21	134830	584	**0.00477**	0.00433	0.00043593
66	*dumat′*	1818	1	170605	4	**0.00055**	0.00002	0.000526609
67	*ožidat′*	17055	19	25570	12	**0.00111**	0.00047	0.000644743
68	*govorit′*	5820	5	321257	8	**0.00086**	0.00002	0.000834204
69	*stojat′*	13502	15	52706	2	**0.00111**	0.00004	0.001073
70	*ugrožat′*	2520	3	6499	0	**0.00119**	0.00000	0.001190476
71	*obtjagivat′*	552	4	168	1	**0.00725**	0.00595	0.001293996
72	*obsuždat′*	4260	8	8461	1	**0.00188**	0.00012	0.001759745

Table D.1 – continued on the next page

#	Lemma	FreqPresp	*očen′* FreqPresp	FreqV	*očen′* FreqV	*očen′* RatioPresp	*očen′* RatioV	Idx
73	*čitat′*	3000	6	39087	4	**0.00200**	0.00010	0.001897664
74	*znat′*	9035	20	379397	32	**0.00221**	0.00008	0.002129269
75	*oblegat′*	1402	11	419	2	**0.00785**	0.00477	0.003072665
76	*pit′*	800	3	20403	4	**0.00375**	0.00020	0.00355395
77	*zapominat′sja*	6543	52	3062	13	**0.00795**	0.00425	0.003701834
78	*objazyvat′*	1122	6	4888	7	**0.00535**	0.00143	0.003915515
79	*zavoraživat′*	1456	14	2020	8	**0.00962**	0.00396	0.005654989
80	*zaxvatyvat′*	9520	120	5480	37	**0.01261**	0.00675	0.005853217
81	*xrustet′*	2933	18	456	0	**0.00614**	0.00000	0.006137061
82	*ščadit′*	2862	30	465	2	**0.01048**	0.00430	0.006181105
83	*vydavat′sja*	784	5	22153	1	**0.00638**	0.00005	0.00633241
84	*poseščat′*	4717	37	15532	6	**0.00784**	0.00039	0.007457669
85	*raspolagat′*	1523	18	25510	99	**0.01182**	0.00388	0.007937948
86	*verovat′*	1894	17	1320	0	**0.00898**	0.00000	0.008975713
87	*ranit′*	72	2	617	7	**0.02778**	0.01135	0.016432559
88	*uvlekat′sja*	620	18	5553	50	**0.02903**	0.00900	0.020028116
89	*značit′*	1653	116	68011	2	**0.07018**	0.00003	0.070146032
90	*ustrašat′*	849	2	0	0	**0.00236**	0.00000	0.002355713

Table D.1: The complete distribution of participial/verbal word forms and the *očen′* constructions with which they combine.

2 Semantic classes

Lemma	Semantic class
nravit′sja	psych
volnovat′sja	psych
ljubit′	psych
radovat′	psych
uvažat′	psych
smaxivat′	perc
pereživat′	psych
cenit′	psych
bodrit′	changest
obodrjat′	psych
vdoxnovljat′	psych
intrigovat′	psych
motivirovat′	psych
volnovat′	psych
osvežat′	changest
zatjagivat′	changest
bojat′sja	psych
interesovat′	psych
somnevat′sja	ment

Table D.2 – continued on the next page

Lemma	**Semantic class**
napominat′	ment
rekomendovat′	speech
pomogat′	psych
interesovat′sja	ment
uspokaivat′	psych
pugat′	psych
obnadeživat′	psych
mešat′	psych
nuždat′sja	be:exist
šokirovat′	psych
privlekat′	psych
razbirat′sja	ment
počitat′	psych
stradat′	psych
vpečatljat′	psych
različat′sja	perc
žaždat′	psych
podxodit′	ment
stimulirovat′	psych/changest
vlijat′	psych/changest
otličat′sja	perc
razvivat′	changest
sposobstvovat′	changest
podderživat′	contact
zaviset′	be:exist
želat′	psych
umet′	ment
stremit′sja	ment
idti	perc
potrjasat′	psych
goret′	light
sootvetstvovat′	ment
vladet′	poss
uznavat′	ment
rabotat′	impact:creat/be:creat
pol′zovat′sja	poss
obitat′	be:exist/loc
upravl jat′	changest
trebovat′	speech
oxranjat′	poss
proxodit′	move
stoit′	poss
vyzyvat′	speech
blestet′	light
vozbuždat′	psych
ponimat′	ment
dumat′	ment
ožidat′	ment
govorit′	speech

Table D.2 – continued on the next page

Lemma	Semantic class
stoit′[1]	poss
ugrožat′	speech
obtjag ivat′	contact
obsuždat′	speech
čitat′	perc
znat′	ment
oblegat′	contact
pit′	physiol
zapominat′sja	ment
objazyvat′	psych
zavoraživat′	psych
zaxvatyvat′	psych
xrustet′	sound
ščadit′	psych
vydavat′sja	perc
poseščat′	move
raspolagat′	psych
verovat′	ment
ranit′	impact
uvlekat′sja	psych
značit′	ment
ustrašat′	psych

Table D.2: The list of the lemmas annotated with the tags indicating semantic classes.

3 *Očen′* constructions with the highest ratio

Lemma	FreqPresp	FreqPresp *očen′*	FreqV	FreqV *očen′*	RatioPresp *očen′*	RatioV *očen′*
ljubit′	15316	237	164104	16210	0.01547	0.09878
napominat′	12520	208	46190	1603	0.01661	0.03470
podxodit′	19977	79	133325	1314	0.00395	0.00986
idti	15381	5	235019	258	0.00033	0.00110
znat′	9035	20	379397	32	0.00221	0.00008
zapominat′sja	6543	52	3062	13	0.00795	0.00425
zaxvatyvat′	9520	120	5480	37	0.0 1261	0.00675
značit′	1653	116	68011	2	0.07018	0.00003

Table D.3: A sample of the eight constructions of *očen′* + PRESP and *očen′* + PRS V that had the highest frequency among other constructions in the list presented in Table D.1.

[1] The initial lemma *stojat′* ‘stand’ that was found in the corpus was incorrectly assigned to the word forms, and was replaced with *stoit′* ‘cost, be worth’.

Appendix E

Statistical analysis: adverbial *očen′* construction

```
ochen <- read.csv("D:/ochen-stat/sem-ratio.csv”)
ochen$semrole <- factor(ochen$semrole)

#currently 16 levels =>
summary(ochen$semrole)
be:exist      be:exist/loc      changest
2           1           6
contact        impact impact:creat/be:creat
3           1           1
light         ment         move
2          16           2
perc        physiol         poss
6           1           5
psych    psych/changest         sound
35           2           1
speech
6
> ochen$semrole <- fct_lump(ochen$semrole, 4, ties.method = "average")
> summary(ochen$semrole)
changest   ment   perc  psych  speech  Other
6    16    6    35    6    21

> ochen$ratioptcp <- squish(ochen$ratioptcp, quantile(ochen$ratioptcp,c(.05,.95)))
> ochen$ratiov <- squish(ochen$ratiov, quantile(ochen$ratioptcp,c(.05,.95)))

> summary(ochen$ratiov)
Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
0.000000 0.000275 0.004315 0.009714 0.018523 0.027949
> summary(ochen$ratioptcp)
Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
0.000000 0.000205 0.001685 0.005111 0.006332 0. 028088

#Standard deviation scores
> sd(ochen$ratiov)
[1] 0.01083359
> sd(ochen$ratioptcp)
[1] 0.007710632
```

```
#Skewness estimation
> skewness(ochen$ratiov)
[1] 0.7519275
> skewness(ochen$ratioptcp)
[1] 1.90283 9

#Mean scores for each level of semrole (sorted by ratiov and ratioptcp)
> mean_ptcp = aggregate(ratioptcp~semrole, data=ochen, FUN=mean)
> mean_ptcp
semrole  ratioptcp
1 changest 0.001393333
2   ment 0.005173625
3   perc 0.002075000
4  psych 0.008097200
5  speech 0.000735000
6  Other 0.003267143
> mean_v = aggregate(ratiov~semrole, data=ochen, FUN=mean)
> mean_v
semrole   ratiov
1 changest 0.01510 1667
2   ment 0.006533125
3   perc 0.014208333
4  psych 0.031533429
5  speech 0.002895000
6  Other 0.002370952

#Boxplots of the ratio distributions
>fun_mean <- function(x){return(round(data.frame(y
↪  =mean(x),label=mean(x,na.rm=T)),digit=5))}

> ggboxplot(data=ochen, x = "semrole", y = "ratioptcp", color="black", fill = "lightblue1",
↪  ylab = "Ratio of PRS
+ PTCP constructions", xlab = "Sematic Roles")+stat_summary(fun.y=mean,geom ="point",size =
↪  1.5,color="red")+theme_bw()+ stat_summary(fun.data = fun_mean, geom="text",
↪  color="red", size =3.7, vjust=-0.5)

> ggboxplot(data=ochen, x = "semrole", y = "ratiov", fill="lightblue1", color="black", ylab
↪  = "Ratio of ochen PRS V constructions", xlab = "Sematic
↪  roles")+stat_summary(fun.y=mean,geom                ="point",size=2,color
↪  ="red")+theme_bw()+ stat_summary(fun.data = fun_mean, geom="text", size =3.2,
↪  vjust=-0.5)

# independent 2-group Mann-Whitney U Test

> wilcox.test(ochen$ratioptcp,ochen$ratiov)

Wilcoxon rank sum test with continuity correction

data: ochen$ratioptcp and ochen$ratiov
W = 3313.5, p-value = 0.03506
alternative hypothesis: true location shift is not equal to 0

# Kruskal-Wallis rank sum test
> ptcp.krusk.t <- kruskal.test(ochen$ratioptcp~ochen$semrole)
> ptcp.krusk.t

Kruskal-Wallis rank sum test
```

```
data: ochen$ratioptcp by ochen$semrole
Kruskal-Wallis chi-squared = 11.37, df = 5, p-value = 0.04452

> v.krusk.t <- kruskal.test(ochen$ratiov~ochen$semrole)
> v.krusk.t

Kruskal-Wallis rank sum test

data: ochen$ratiov by ochen$semrole
Kruskal-Wallis chi-squared = 32.732, df = 5, p-value = 4.255e-06

# Dunn's test
> ptcp.dunn.t <- ochen %>% dunn_test(ratioptcp ~ semrole, p.adjust.method = "bonferroni")
> ptcp.dunn.t
# A tibble: 15 x 9
.y.    group1  group2  n1  n2 statistic   p p.adj p.adj.signif
* <chr>   <chr>  <chr> <int> <int>   < dbl> <dbl> <dbl> <chr>
1 ratioptcp changest ment    6  16   1.50 0.133 1   ns
2 ratioptcp changest perc    6   6   0.702 0.483 1   ns
3 ratioptcp changest psych   6  35   2.11 0.0350 0.526 ns
4 ratioptcp changest speech   6   6   0   1   1   ns
5 ratioptcp changest Other   6  21   0.503 0.615 1   ns
6 ratioptcp ment   perc   16   6  -0.654 0.513 1   ns
7 ratioptcp ment   psych   16  35   0.705 0.481 1   ns
8 ratioptcp ment   speech  16   6  -1.50 0.133 1   ns
9 ratioptcp ment   Other   16  21  -1.46 0.143 1   ns
10 ratioptcp perc   psych   6  35   1.19 0.234 1   ns
11 ratioptcp perc   speech   6   6   -0.702 0.483 1   ns
12 ratioptcp perc   Other   6  21  -0.372 0.710 1   ns
13 ratioptcp psych  speech  35   6  -2.11 0.0350 0.526 ns
14 ratioptcp psych  Other   35  21  -2.53 0.0114 0.171 ns
15 ratioptcp speech  Other   6  21   0.503 0.615 1   ns


> v.dunn.t <- ochen %>% dunn_test(ratiov ~ semrole, p.adjust.method = "bonferroni")
> v.dunn.t
# A tibble: 15 x 9
.y.  group1  group2  n1  n2 statistic     p   p.adj p.adj.signif
* <chr> <chr>  <chr> <int> <int>   <dbl>   <dbl>   <dbl> <chr>
1 ratiov changest ment    6  16  -1.06 0.287    1     ns
2 ratiov changest perc    6   6  -0.638 0.524   1     ns
3 ratiov changest psych   6  35   1.11 0.267    1     ns
4 ratiov changest speech   6   6  -1.95 0.0509   0.763   ns
5 ratiov changest Other   6  21  -1.84 0.0655   0.982   ns
6 ratiov ment   perc    16   6   0.295 0.768   1     ns
7 ratiov ment   psych   16  35   3.31 0.000921  0.0138  *
8 ratiov ment   speech  16   6  -1.29 0.197   1      ns
9 ratiov ment   Other   16  21  -1.03 0.301   1     ns
10 ratiov perc   psych   6  35   1.94 0.0520   0.780   ns
11 ratiov perc   speech   6    6  -1.31 0.189   1     ns
12 ratiov perc   Other   6  21  -1.05 0.295   1     ns
13 ratiov psych  speech  35   6  -3.66 0.000251  0.00377  **
14 ratiov psych  Other   35  21  -4.87 0.00000114 0.0000171 ****
15 ratiov speech  Other   6  21   0.593 0.553   1     ns
```

Appendix F

Distribution analysis

Present active	**Past active**	**Present passive**	**Past passive**
-ušč-/-jušč-	*-vš-*	*-em-/-om-*	*-nn-*
nesuščij pojuščij	pisavšij darivšij	izučaemyj vedomyj	zvannyj izbrannyj izmotannyj
-ašč-/-jašč-	*-š-*	*-im-*	*-enn-*
ležaščij strojaščij	zabredšij otsvetšij umeršij	slyšimyj gonimyj	uvlečënnyj ušiblennyj nošennyj
			-t-
			kinutyj zavernutyj

Table F.1: Suffixes for the full forms of participles; that is, present/past active/passive forms.

Present/passive	**Past/passive**
-en-/-n-	*-em-/-om-/-im-*
obižen-a-o-y narisovan-a-o-y	čitaem-a-o-y vedom-a-o-y terpim-a-o-y
	-t-
	bit-a-o-y

Table F.2: Suffixes for the short forms of participles; that is, present/past active/passive forms.

<table>
<tr><th>Passive</th><th>Active</th></tr>
<tr><td>yj|ogo|omu|ym|om|oe|ogo|omu
|aja|oj|uju|oju|ye|yx| ym|ymi</td><td>ij|ego|emu|ij|im|em|ee
|aja|ej|uju|eju|ie|ix|imi</td></tr>
</table>

Table F.3: Forms of inflection in the declension of passive and active participles.

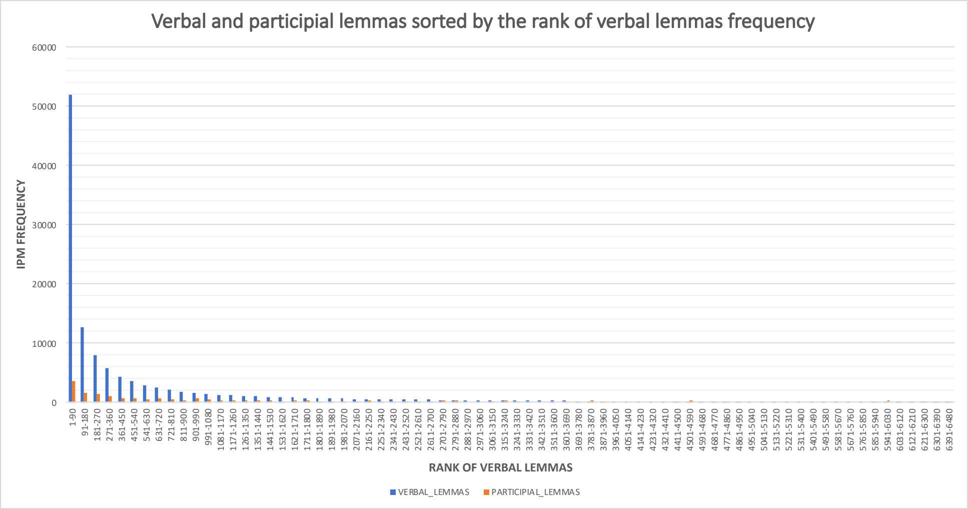


Figure F.1: Distribution of verbal and participial lemmas depending on the frequency ranks of verbal lemmas.

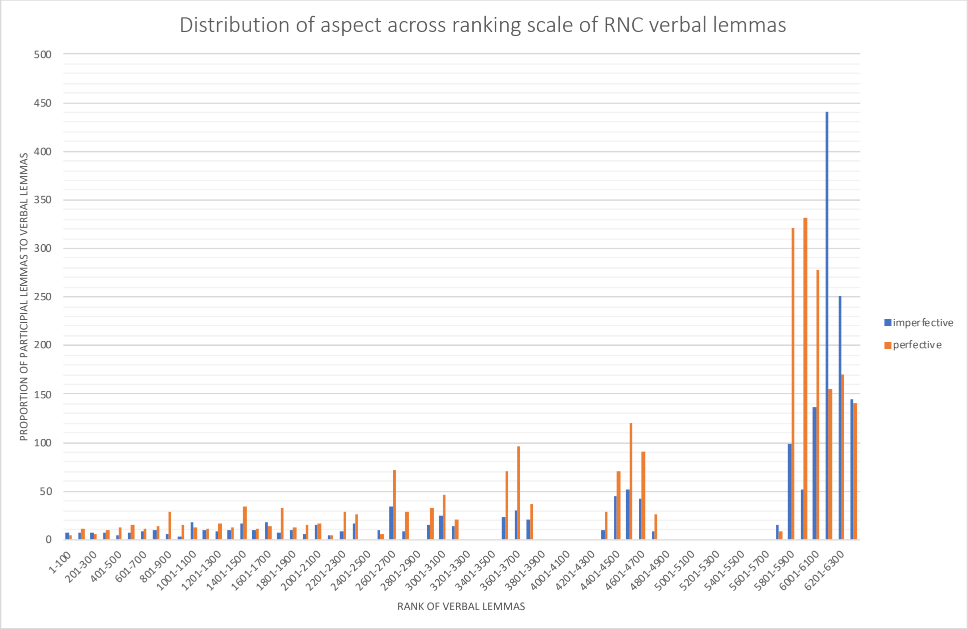


Figure F.2: Ratio of imperfective and perfective participial lemmas to verbal lemmas distributed across the ranks of the verbal lemmas.

Verbal lemma	**Rank**	**Trans Asp**
xotet′ ‘want’	50	INTR IPFV
sprosit′ ‘ask’	39	TR PFV
xotet′sja ‘feel like’	50	I NTR IPFV
smoč′ ‘manage’	71	INTR PFV
vesit′ ‘weigh’	78	TR PFV
pomoč′ ‘help’	93	INTR PFV
prijtis′ ‘have to’	99	INTR PFV

Table F.4: Verbal lemmas with no corresponding participles within the range of 1–100.

Verbal lemma	**Rank**	**Trans Asp**
poležat′ lie for a while’	3054	INTR PFV
pomaxat′ ‘dangle’	3055	INTR PFV
pereska zyvat′ ‘retell’	3034	TR IPFV
otodvinut′sja‘move aside’	3035	INTR PFV
ladit′ ‘get along’	2952	INTR IPFV
snjat′sja ‘act in a film’	2946	INTR PFV

(a) Interval 2881–3060.

Verbal lemma	**Rank**	**Trans Asp**
načisljat′ ‘sew on’	33 32	TR IPFV
zavalivat′ ‘lumber’	3333	TR IPFV
zavyt′ ‘howl’	3482	INTR IPFV
klevat′‘peck’	3483	TR IPFV
umaljat′ ‘plead’	3520	TR IPFV
tresnut′ ‘creak’	3560	INTR PFV

(b) Interval 3871–4320.

Verbal lemma	**Rank**	**Trans Asp**
natolknut′sja ‘run into’	3945	INTR PFV
pomërznut′ ‘freeze for a while’	3946	INTR PFV
priš vcemit′ ‘squeeze’	4066	TR PFV
rasčistit′‘clear out’	4172	TR PFV
pridirat′ sja‘nag’	4173	INTR IPFV
portit ’sja ‘spoil’	4315	INTR IPFV

(c) Interval 3871–4320.

Table F.5: The list of verbal lemmas without corresponding participles within the mid-frequency range.

Verbal lemma	**Rank**	**Trans Asp**
vvalit′sja ‘burst in ’	4331	INTR PFV
kipjatit′ ‘boil’	4332	TR IPFV
navredit′ ‘harm’	4349	INTR PFV
izumljat′ ‘amaze’	4377	TR IPFV
zamj at′sja′ ‘falter’	4394	INTR PFV
soveršenstvovat′ ‘improve’	4395	TR IPFV

(a) Interval 4321–4410.

Verbal lemma	**Rank**	**Trans Asp**
oblokotit′sja ‘lean elbow’	4963	INTR PFV
utomljat′ ‘tire’	49 64	TR IPFV
čertyxat′sja ‘swear’	5336	INTR IPFV
otključitś ‘turn off’	5330	INTR PFV
zaščitit′sja ‘defend’	3520	INTR PFV
razbivat′ ‘break/split’	5783	TR IPFV

(b) Interval 4951–5850.

Table F.6: The list of verbal lemmas without corresponding participles within the low-frequency range.

Appendix G

Statistical analysis: frequency distributions

1 Binary logistic regression models

1.1 Model 1: frequency, ratio

```
freq <- read.table("/Users/upe007/Box
↪  Sync/diss/diss_exp/ch4.3/glm/analys/dataset/freq.ratio.ambig.csv",header=TRUE, sep=",")

nrow(freq)
[1] 3336

freq$ambiguity <- factor(freq$ambiguity)
freq$ipmVerb <- log1p(freq$ipmVerb)
freq$ipmVerb1 <- squish(freq$ipmVerb, quantile(freq$ipmVerb, c(.05, .95)))
freq$ratioPtcp1 <- squish(freq$ratioPtcp, quantile(freq$ratioPtcp, c(.05, .95)))

> summary(freq)
ipmVerb     ratioPtcp     ambiguity  ipmVerb1    ratioPtcp1
Min.  :0.3289  Min.  : 0.00026  0:2942  Min.  :0.3289  Min.  :0.03738
1st Qu.:0.4203  1st Qu.: 0.15385  1: 394  1st Qu.:0.4203  1st Qu.:0.15385
Median :0.5351  Median :0.40122       Median :0.5351  Median :0.40122
Mean  :0.5206  Mean  :1.03036       Mean  :0.5196  Mean  :0.77339
3rd Qu.:0.6189  3rd Qu.:1.00000       3rd Qu.:0.6189  3rd Qu.:1.00000
Max.  :0.7887  Max.  :122.00000       Max.  :0.6951  Max.  :3.50000

fit.freq = glm(ambiguity ~ ratioPtcp1 + ipmVerb1 + ratioPtcp1*ipmVerb1, family =
↪  binomial(link = "logit"),data=freq)
summary(fit.freq)

Call:
glm(formula = ambiguity ~ ratioPtcp1 + ipmVerb1 + ratioPtcp1 *
ipmVerb1, family = binomial(link = "logit"), data = freq)

Deviance Residuals:
Min     1Q  Median    3Q    Max
-2.9467 -0.4190 -0.2777 -0.2086  2.3131
```

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.39381  0.23667 -22.790 < 2e-16 ***
ratioPtcp1       1.14055  0.10279 11.095 < 2e-16 ***
ipmVerb1        0.79622  0.07015 11.351 < 2e-16 ***
ratioPtcp1:ipmVerb1 0.32387  0.06837  4.737 2.17e-06 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2422.8 on 3335 degrees of freedom
Residual deviance: 1810.3 on 3332 degrees of freedom
AIC: 1818.3

Number of Fisher Scoring iterations: 6

#-----------------------------
rsq.kl(fit.freq)
[1] 0.2528295

#-----------------Testing model fit-------------------
fit.freq.null = glm(ambiguity ~ 1, family=binomial(link="logit"),data = freq)
anova(fit.freq.null,fit.freq,test = "LRT")
Analysis of Deviance Table

Model 1: ambiguity ~ 1
Model 2: ambiguity ~ ratioPtcp1 + ipmVerb1 + ratioPtcp1 * ipmVerb1
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1   3335   2422.8
2   3332   1810.3 3  612.56 < 2.2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.2 Model 2: tense, voice, frequency

```
tense.voice <- read.table("/Users/upe007/Box
↪  Sync/diss/diss_exp/ch4.3/glm/analys/dataset/tense.voice.ipm.ratio.ambg.csv",
↪  header=TRUE, sep=",")

nrow(tense.voice)
[1] 3384

tense.voice$ambiguity <- factor(tense.voice$ambiguity)
tense.voice$ipmVerb1 <- squish(tense.voice$ipmVerb, quantile(tense.voice$ipmVerb, c(.05,
↪  .95)))
tense.voice$ratioPtcp1 <- squish(tense.voice$ratioPtcp, quantile(tense.voice$ratioPtcp,
↪  c(.05, .95)))
tense.voice$ipmVerb1 <- log1p(tense.voice$ipmVerb1)

summary(tense.voice)
tense    voice     ipmVerb     ratioPtcp    ambiguity  ipmVerb1    ratioPtcp1
praes:1212  act :1909  Min.  :  0.80  Min.  : 0.00035  0:2918  Min.  :0.5878  Min.  :0.06701
praet:2172  pass:1475  1st Qu.:  1.70  1st Qu.: 0.27237  1: 466  1st Qu.:0.9933  1st
↪  Qu.:0.27237
Median :  3.40  Median : 0.58621       Median :1.4816  Median :0.58621
```

```
Mean  : 17.51  Mean  : 0.79580       Mean  :1.7455  Mean  :0.64512
3rd Qu.: 10.10  3rd Qu.: 1.00000       3rd Qu.:2.4069  3rd Qu.:1.00000
Max.  :5403.40  Max.  :61.00000       Max.  :4.1043  Max.  :1.50000

fit.tense.voice=glm(ambiguity ~
↪  tense+voice+tense*scale(ipmVerb1,scale=F)+voice*scale(ipmVerb1,scale=F)+tense*voice,
↪  family=binomial(link="logit"),contrasts=list(tense=contr.sum(2),voice=contr.sum(2)),
↪  data=tense.voice)

summary(fit.tense.voice)

Call:
glm(formula = ambiguity ~ tense + voice + tense * scale(ipmVerb1,
scale = F) + voice * scale(ipmVerb1, scale = F) + tense *
voice, family = binomial(link = "logit"), data = tense.voice,
contrasts = list(tense = contr.sum(2), voice = contr.sum(2)))

Deviance Residuals:
Min    1Q  Median    3Q   Max
-0.7852 -0.5768 -0.5171 -0.4326  2.8386

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.43061  0.18376 -13.227 < 2e-16 ***
tense1              -0.52464  0.18272 -2.871 0.00409 **
voice1               0.42853  0.18332  2.338 0.01941 *
scale(ipmVerb1, scale = F)     0.26615  0.05487  4.851 1.23e-06 ***
tense1:scale(ipmVerb1, scale = F) -0.01772  0.05851 -0.303 0.76202
voice1:scale(ipmVerb1, scale = F) 0.02050  0.06103  0.336 0.73692
tense1:voice1           0.69463  0.18398  3.776 0.00016 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2712.5 on 3383 degrees of freedom
Residual deviance: 2645.6 on 3377 degrees of freedom
AIC: 2659.6

Number of Fisher Scoring iterations: 6

rsq.kl(fit.tense.voice)
[1] 0.02463493

#------------------Testing model fit---------------------------
fit.tense.voice.null = glm(ambiguity ~ 1, family=binomial(link="logit"),data = tense.voice)
anova(fit.tense.voice.null,fit.tense.voice,test = "LRT")
Analysis of Deviance Table

Model 1: ambiguity ~ 1
Model 2: ambiguity ~ tense + voice + tense * scale(ipmVerb1, scale = F) +
voice * scale(ipmVerb1, scale = F) + tense * voice
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1   3383   2712.5
2   3377   2645.7 6  66.822 1.831e-12 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.3 Model 3: aspect, transitivity, frequency

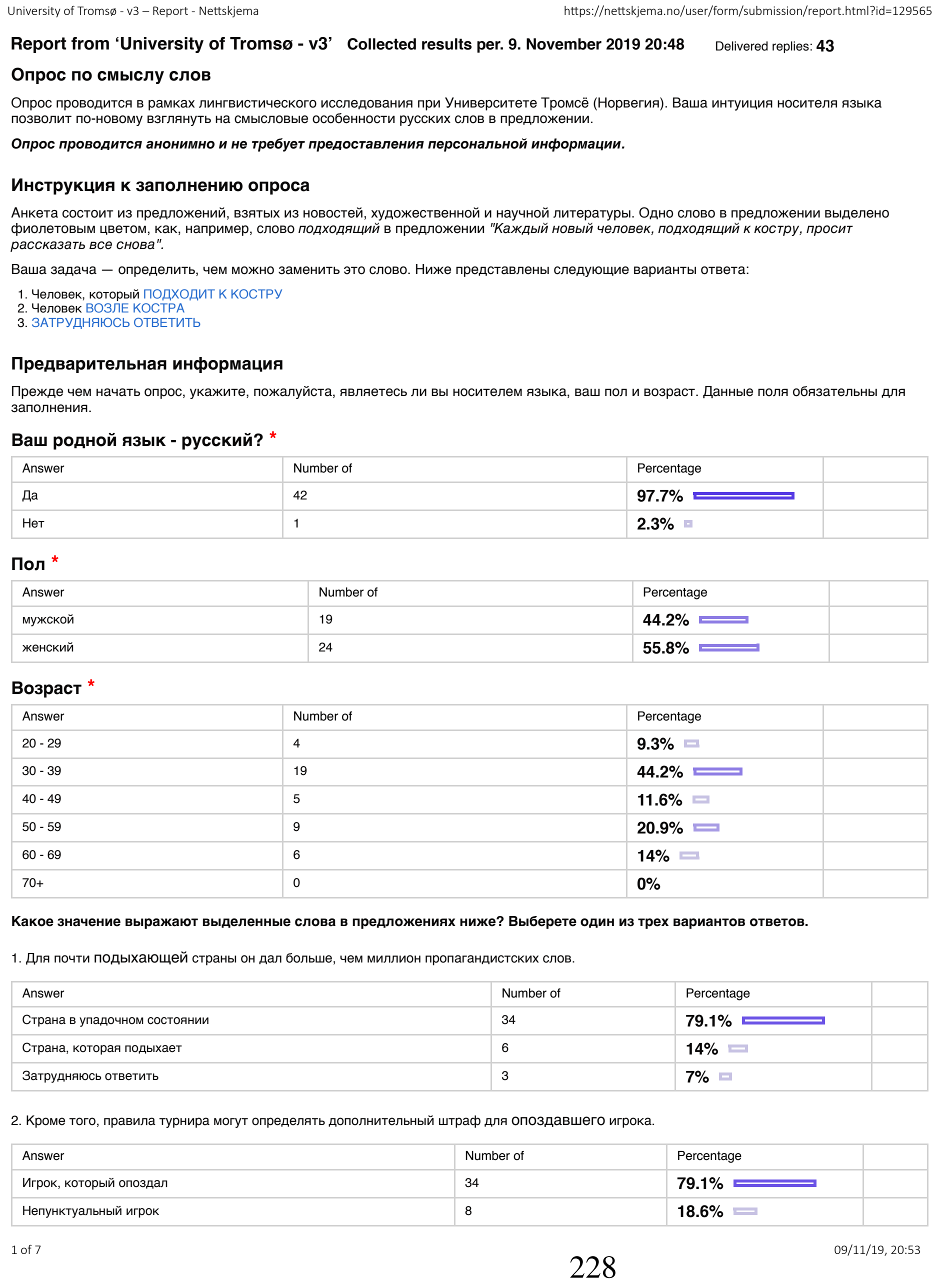
```
> trans.asp <- read.table("/Users/upe007/Box
↪  Sync/diss/diss_exp/exp4/glm/analys/dataset/trans.asp.ipm.ratio.ambg.csv",header=TRUE,
↪  sep=",")

nrow(trans)
[1] 2678

trans.asp$ambiguity <- factor(trans.asp$ambiguity)
trans.asp$ipmVerb <- log1p(trans.asp$ipmVerb)
trans.asp$ipmVerb1 <- squish(trans.asp$ipmVerb, quantile(trans$ipmVerb, c(.05, .95)))
trans.asp$ratioPtcp1 <- squish(trans.asp$ratioPtcp, quantile(trans$ratioPtcp, c(.05, .95)))

summary(trans.asp)
lemma   transitivity aspect   ipmVerb     ratioPtcp     ambiguity  ipmVerb1    ratioPtcp1
быть   :  3  intr: 677  ipf:1303  Min.  :0.5878  Min.  : 0.00028  0:2304  Min.  :0.5878  Min.  :0.03689
выходить :  3  tran:2001  pf :1375  1st Qu.:0.9933  1st Qu.: 0.16667  1: 374  1st Qu.:0.9933  1st
↪  Qu.:0.16667
засыпать :  3              Median :1.7918  Median : 0.50000        Median :1.7918  Median :0.50000
переходить:  3              Mean  :2.0763  Mean  : 1.37269        Mean  :2.0394  Mean  :0.91959
проходить :  3              3rd Qu.:2.8792  3rd Qu.: 1.00000        3rd Qu.:2.8792  3rd Qu.:1.00000
смотреть :  3              Max.  :9.3896  Max.  :163.00000        Max.  :4.7587  Max.  :4.25197
(Other)  :2660

fit.trans.asp=glm(ambiguity ~
↪  transitivity+aspect+transitivity*scale(ipmVerb1,scale=F)+aspect*scale(ipmVerb1,scale=F)
↪  +transitivity*aspect,family=binomial(link="logit"),contrasts=list(transitivity=contr.sum(2),
↪  aspect=contr.sum(2)), data=trans.asp)

summary(fit.trans.asp)

Call:
glm(formula = ambiguity ~ transitivity + aspect + transitivity *
scale(ipmVerb1, scale = F) + aspect * scale(ipmVerb1, scale = F) +
transitivity * aspect, family = binomial(link = "logit"),
data = trans.asp, contrasts = list(transitivity = contr.sum(2),
aspect = contr.sum(2)))

Deviance Residuals:
Min    1Q  Median    3Q   Max
-0.7190 -0.6073 -0.5148 -0.4348  2.2197

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.967331  0.075742 -25.974 < 2e-16 ***
transitivity1             -0.085926  0.075373 -1.140 0.254283
aspect1                 -0.007605  0.075011 -0.101 0.919244
scale(ipmVerb1, scale = F)         0.072895  0.050262  1.450 0.146975
transitivity1:scale(ipmVerb1, scale = F) 0.029020  0.052088  0.557 0.577438
aspect1:scale(ipmVerb1, scale = F)     0.111341  0.048265  2.307 0.021061 *
transitivity1:aspect1            0.260211  0.074880  3.475 0.000511 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2165.6 on 2677 degrees of freedom
```

```
Residual deviance: 2134.3 on 2671 degrees of freedom
AIC: 2148.3

Number of Fisher Scoring iterations: 4

rsq.kl(fit.trans.asp)
[1] 0.01449283

#----------------------Testing model fit------------------------
fit.trans.asp.null = glm(ambiguity ~ 1, family=binomial(link="logit"),data = trans.asp)
anova(fit.trans.asp.null,fit.trans.asp,test = "LRT")
Analysis of Deviance Table

Model 1: ambiguity ~ 1
Model 2: ambiguity ~ transitivity + aspect + transitivity * scale(ipmVerb1,
scale = F) + aspect * scale(ipmVerb1, scale = F) + transitivity *
aspect
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1   2677   2165.6
2   2671   2134.3 6  31.386 2.139e-05 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix H

Survey design

1 Survey report



University of Tromsø - v3 – Report - Nettskjema https://nettskjema.no/user/form/submission/report.html?id=129565

Report from 'University of Tromsø - v3' Collected results per. 9. November 2019 20:48 Delivered replies: **43**

Опрос по смыслу слов

Опрос проводится в рамках лингвистического исследования при Университете Тромсё (Норвегия). Ваша интуиция носителя языка позволит по-новому взглянуть на смысловые особенности русских слов в предложении.

Опрос проводится анонимно и не требует предоставления персональной информации.

Инструкция к заполнению опроса

Анкета состоит из предложений, взятых из новостей, художественной и научной литературы. Одно слово в предложении выделено фиолетовым цветом, как, например, слово *подходящий* в предложении *"Каждый новый человек, подходящий к костру, просит рассказать все снова".*

Ваша задача — определить, чем можно заменить это слово. Ниже представлены следующие варианты ответа:

1. Человек, который ПОДХОДИТ К КОСТРУ
2. Человек ВОЗЛЕ КОСТРА
3. ЗАТРУДНЯЮСЬ ОТВЕТИТЬ

Предварительная информация

Прежде чем начать опрос, укажите, пожалуйста, являетесь ли вы носителем языка, ваш пол и возраст. Данные поля обязательны для заполнения.

Ваш родной язык - русский? *

Answer	Number of	Percentage	
Да	42	**97.7%**	
Нет	1	**2.3%**	

Пол *

Answer	Number of	Percentage	
мужской	19	**44.2%**	
женский	24	**55.8%**	

Возраст *

Answer	Number of	Percentage	
20 - 29	4	**9.3%**	
30 - 39	19	**44.2%**	
40 - 49	5	**11.6%**	
50 - 59	9	**20.9%**	
60 - 69	6	**14%**	
70+	0	**0%**	

Какое значение выражают выделенные слова в предложениях ниже? Выберете один из трех вариантов ответов.

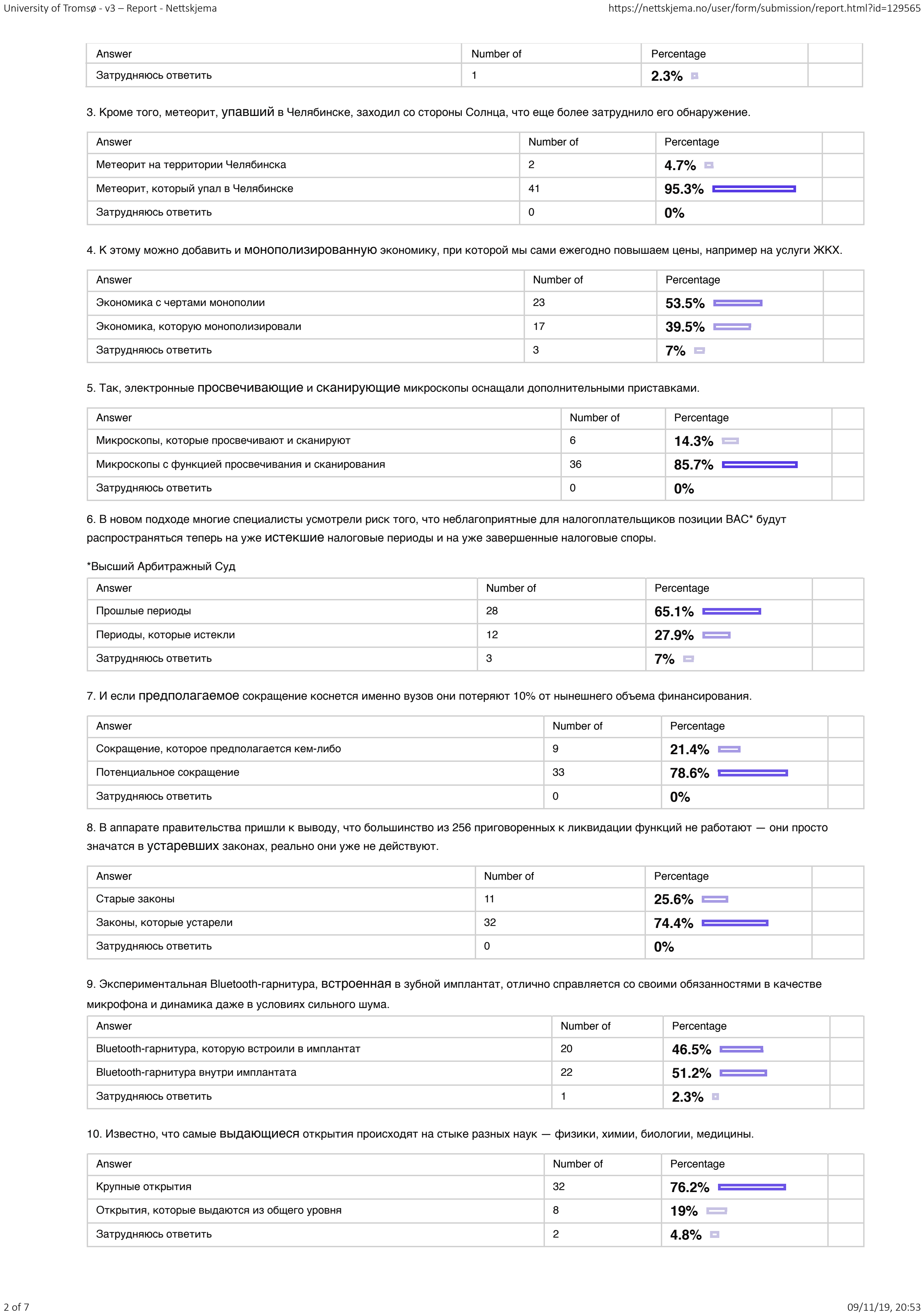
1. Для почти подыхающей страны он дал больше, чем миллион пропагандистских слов.

Answer	Number of	Percentage	
Страна в упадочном состоянии	34	**79.1%**	
Страна, которая подыхает	6	**14%**	
Затрудняюсь ответить	3	**7%**	

2. Кроме того, правила турнира могут определять дополнительный штраф для опоздавшего игрока.

Answer	Number of	Percentage	
Игрок, который опоздал	34	**79.1%**	
Непунктуальный игрок	8	**18.6%**	

1 of 7 09/11/19, 20:53



University of Tromsø - v3 – Report - Nettskjema https://nettskjema.no/user/form/submission/report.html?id=129565

Answer	Number of	Percentage	
Затрудняюсь ответить	1	2.3%	

3. Кроме того, метеорит, упавший в Челябинске, заходил со стороны Солнца, что еще более затруднило его обнаружение.

Answer	Number of	Percentage	
Метеорит на территории Челябинска	2	4.7%	
Метеорит, который упал в Челябинске	41	95.3%	
Затрудняюсь ответить	0	0%	

4. К этому можно добавить и монополизированную экономику, при которой мы сами ежегодно повышаем цены, например на услуги ЖКХ.

Answer	Number of	Percentage	
Экономика с чертами монополии	23	53.5%	
Экономика, которую монополизировали	17	39.5%	
Затрудняюсь ответить	3	7%	

5. Так, электронные просвечивающие и сканирующие микроскопы оснащали дополнительными приставками.

Answer	Number of	Percentage	
Микроскопы, которые просвечивают и сканируют	6	14.3%	
Микроскопы с функцией просвечивания и сканирования	36	85.7%	
Затрудняюсь ответить	0	0%	

6. В новом подходе многие специалисты усмотрели риск того, что неблагоприятные для налогоплательщиков позиции ВАС* будут распространяться теперь на уже истекшие налоговые периоды и на уже завершенные налоговые споры.

*Высший Арбитражный Суд

Answer	Number of	Percentage	
Прошлые периоды	28	65.1%	
Периоды, которые истекли	12	27.9%	
Затрудняюсь ответить	3	7%	

7. И если предполагаемое сокращение коснется именно вузов они потеряют 10% от нынешнего объема финансирования.

Answer	Number of	Percentage	
Сокращение, которое предполагается кем-либо	9	21.4%	
Потенциальное сокращение	33	78.6%	
Затрудняюсь ответить	0	0%	

8. В аппарате правительства пришли к выводу, что большинство из 256 приговоренных к ликвидации функций не работают — они просто значатся в устаревших законах, реально они уже не действуют.

Answer	Number of	Percentage	
Старые законы	11	25.6%	
Законы, которые устарели	32	74.4%	
Затрудняюсь ответить	0	0%	

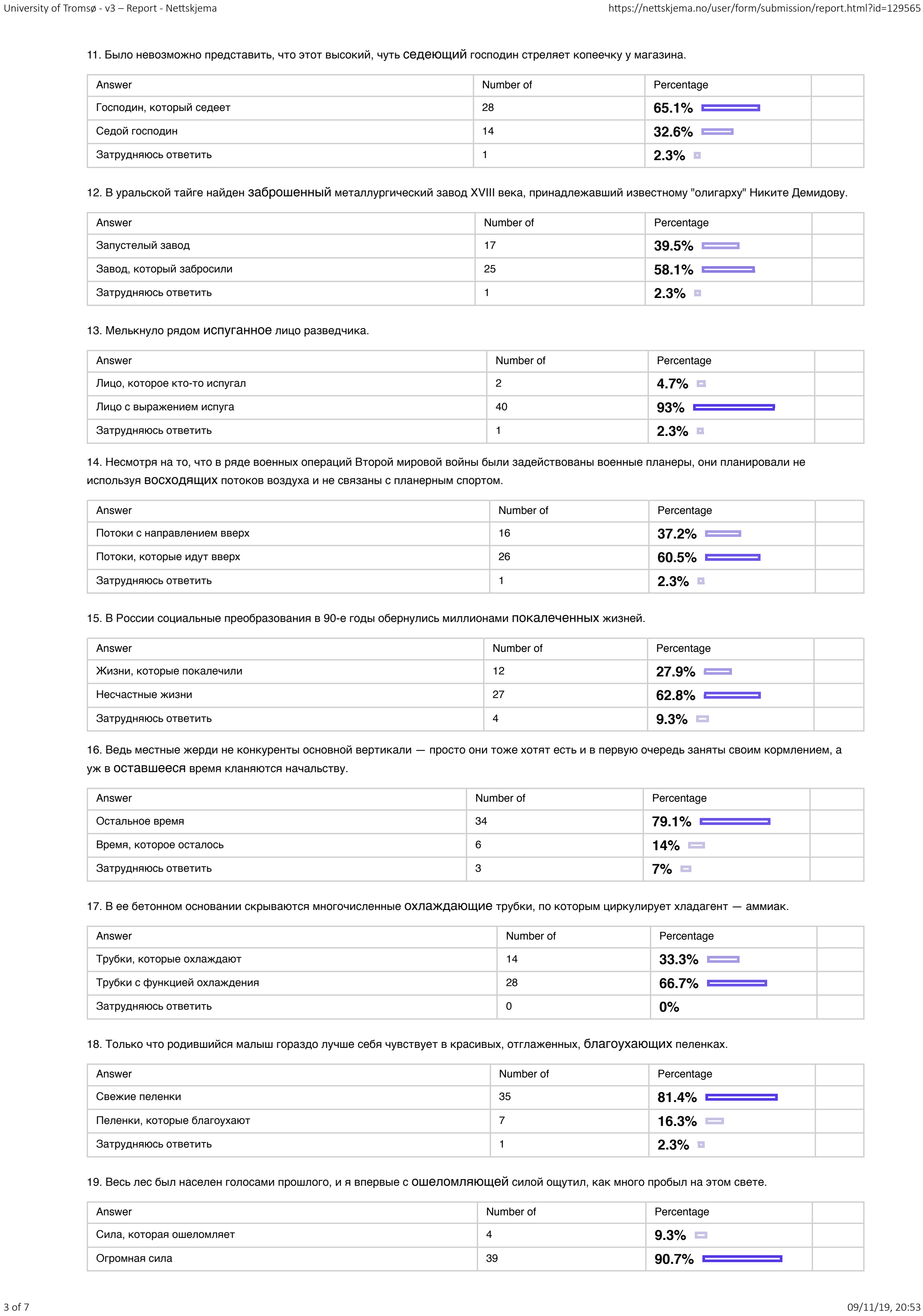
9. Экспериментальная Bluetooth-гарнитура, встроенная в зубной имплантат, отлично справляется со своими обязанностями в качестве микрофона и динамика даже в условиях сильного шума.

Answer	Number of	Percentage	
Bluetooth-гарнитура, которую встроили в имплантат	20	46.5%	
Bluetooth-гарнитура внутри имплантата	22	51.2%	
Затрудняюсь ответить	1	2.3%	

10. Известно, что самые выдающиеся открытия происходят на стыке разных наук — физики, химии, биологии, медицины.

Answer	Number of	Percentage	
Крупные открытия	32	76.2%	
Открытия, которые выдаются из общего уровня	8	19%	
Затрудняюсь ответить	2	4.8%	

2 of 7 09/11/19, 20:53



University of Tromsø - v3 – Report - Nettskjema https://nettskjema.no/user/form/submission/report.html?id=129565

11. Было невозможно представить, что этот высокий, чуть седеющий господин стреляет копеечку у магазина.

Answer	Number of	Percentage	
Господин, который седеет	28	65.1%	
Седой господин	14	32.6%	
Затрудняюсь ответить	1	2.3%	

12. В уральской тайге найден заброшенный металлургический завод XVIII века, принадлежавший известному "олигарху" Никите Демидову.

Answer	Number of	Percentage	
Запустелый завод	17	39.5%	
Завод, который забросили	25	58.1%	
Затрудняюсь ответить	1	2.3%	

13. Мелькнуло рядом испуганное лицо разведчика.

Answer	Number of	Percentage	
Лицо, которое кто-то испугал	2	4.7%	
Лицо с выражением испуга	40	93%	
Затрудняюсь ответить	1	2.3%	

14. Несмотря на то, что в ряде военных операций Второй мировой войны были задействованы военные планеры, они планировали не используя восходящих потоков воздуха и не связаны с планерным спортом.

Answer	Number of	Percentage	
Потоки с направлением вверх	16	37.2%	
Потоки, которые идут вверх	26	60.5%	
Затрудняюсь ответить	1	2.3%	

15. В России социальные преобразования в 90-е годы обернулись миллионами покалеченных жизней.

Answer	Number of	Percentage	
Жизни, которые покалечили	12	27.9%	
Несчастные жизни	27	62.8%	
Затрудняюсь ответить	4	9.3%	

16. Ведь местные жерди не конкуренты основной вертикали — просто они тоже хотят есть и в первую очередь заняты своим кормлением, а уж в оставшееся время кланяются начальству.

Answer	Number of	Percentage	
Остальное время	34	79.1%	
Время, которое осталось	6	14%	
Затрудняюсь ответить	3	7%	

17. В ее бетонном основании скрываются многочисленные охлаждающие трубки, по которым циркулирует хладагент — аммиак.

Answer	Number of	Percentage	
Трубки, которые охлаждают	14	33.3%	
Трубки с функцией охлаждения	28	66.7%	
Затрудняюсь ответить	0	0%	

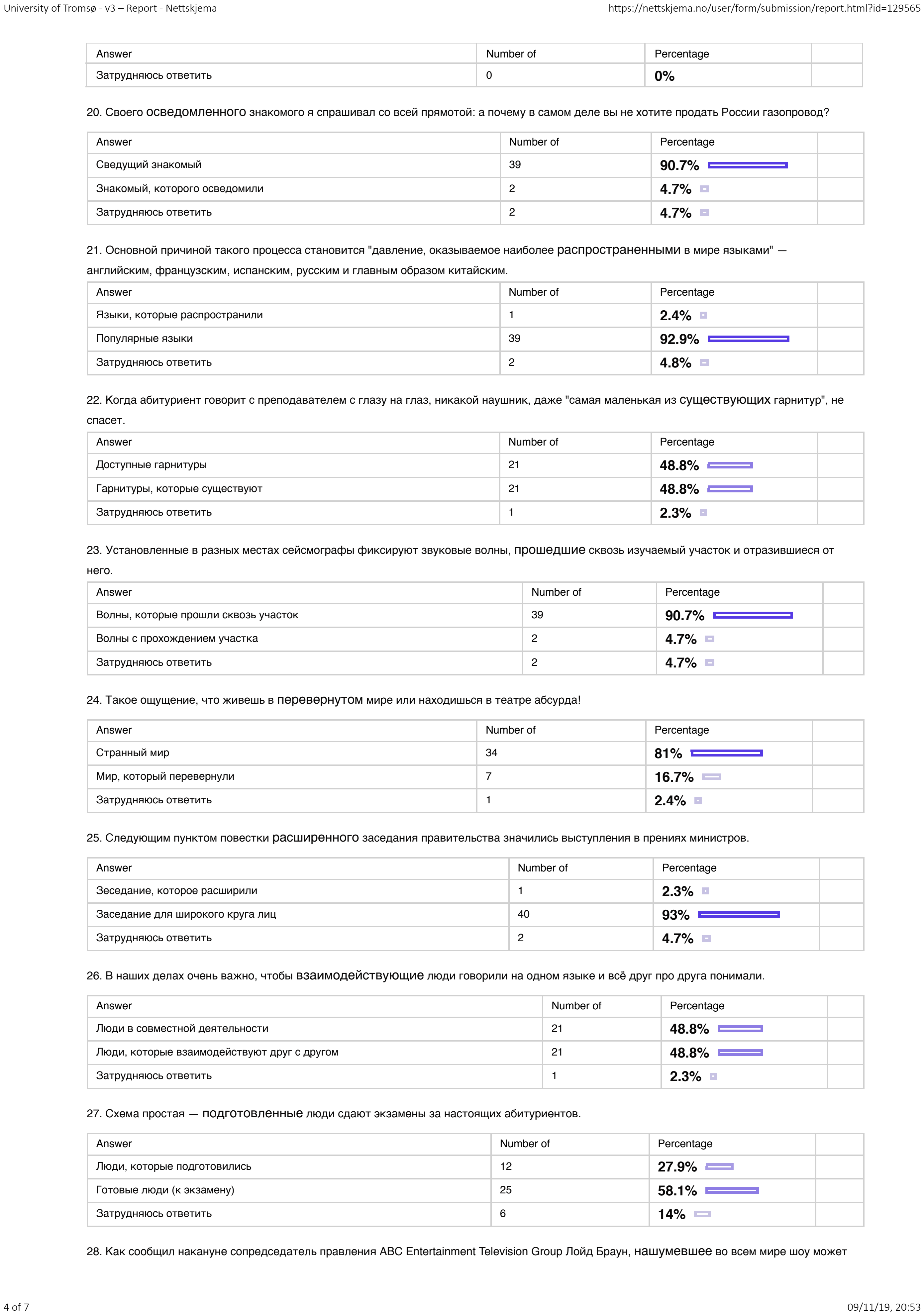
18. Только что родившийся малыш гораздо лучше себя чувствует в красивых, отглаженных, благоухающих пеленках.

Answer	Number of	Percentage	
Свежие пеленки	35	81.4%	
Пеленки, которые благоухают	7	16.3%	
Затрудняюсь ответить	1	2.3%	

19. Весь лес был населен голосами прошлого, и я впервые с ошеломляющей силой ощутил, как много пробыл на этом свете.

Answer	Number of	Percentage	
Сила, которая ошеломляет	4	9.3%	
Огромная сила	39	90.7%	

3 of 7 09/11/19, 20:53



Answer	Number of	Percentage	
Затрудняюсь ответить	0	0%	

20. Своего осведомленного знакомого я спрашивал со всей прямотой: а почему в самом деле вы не хотите продать России газопровод?

Answer	Number of	Percentage	
Сведущий знакомый	39	90.7%	
Знакомый, которого осведомили	2	4.7%	
Затрудняюсь ответить	2	4.7%	

21. Основной причиной такого процесса становится "давление, оказываемое наиболее распространенными в мире языками" — английским, французским, испанским, русским и главным образом китайским.

Answer	Number of	Percentage	
Языки, которые распространили	1	2.4%	
Популярные языки	39	92.9%	
Затрудняюсь ответить	2	4.8%	

22. Когда абитуриент говорит с преподавателем с глазу на глаз, никакой наушник, даже "самая маленькая из существующих гарнитур", не спасет.

Answer	Number of	Percentage	
Доступные гарнитуры	21	48.8%	
Гарнитуры, которые существуют	21	48.8%	
Затрудняюсь ответить	1	2.3%	

23. Установленные в разных местах сейсмографы фиксируют звуковые волны, прошедшие сквозь изучаемый участок и отразившиеся от него.

Answer	Number of	Percentage	
Волны, которые прошли сквозь участок	39	90.7%	
Волны с прохождением участка	2	4.7%	
Затрудняюсь ответить	2	4.7%	

24. Такое ощущение, что живешь в перевернутом мире или находишься в театре абсурда!

Answer	Number of	Percentage	
Странный мир	34	81%	
Мир, который перевернули	7	16.7%	
Затрудняюсь ответить	1	2.4%	

25. Следующим пунктом повестки расширенного заседания правительства значились выступления в прениях министров.

Answer	Number of	Percentage	
Зеседание, которое расширили	1	2.3%	
Заседание для широкого круга лиц	40	93%	
Затрудняюсь ответить	2	4.7%	

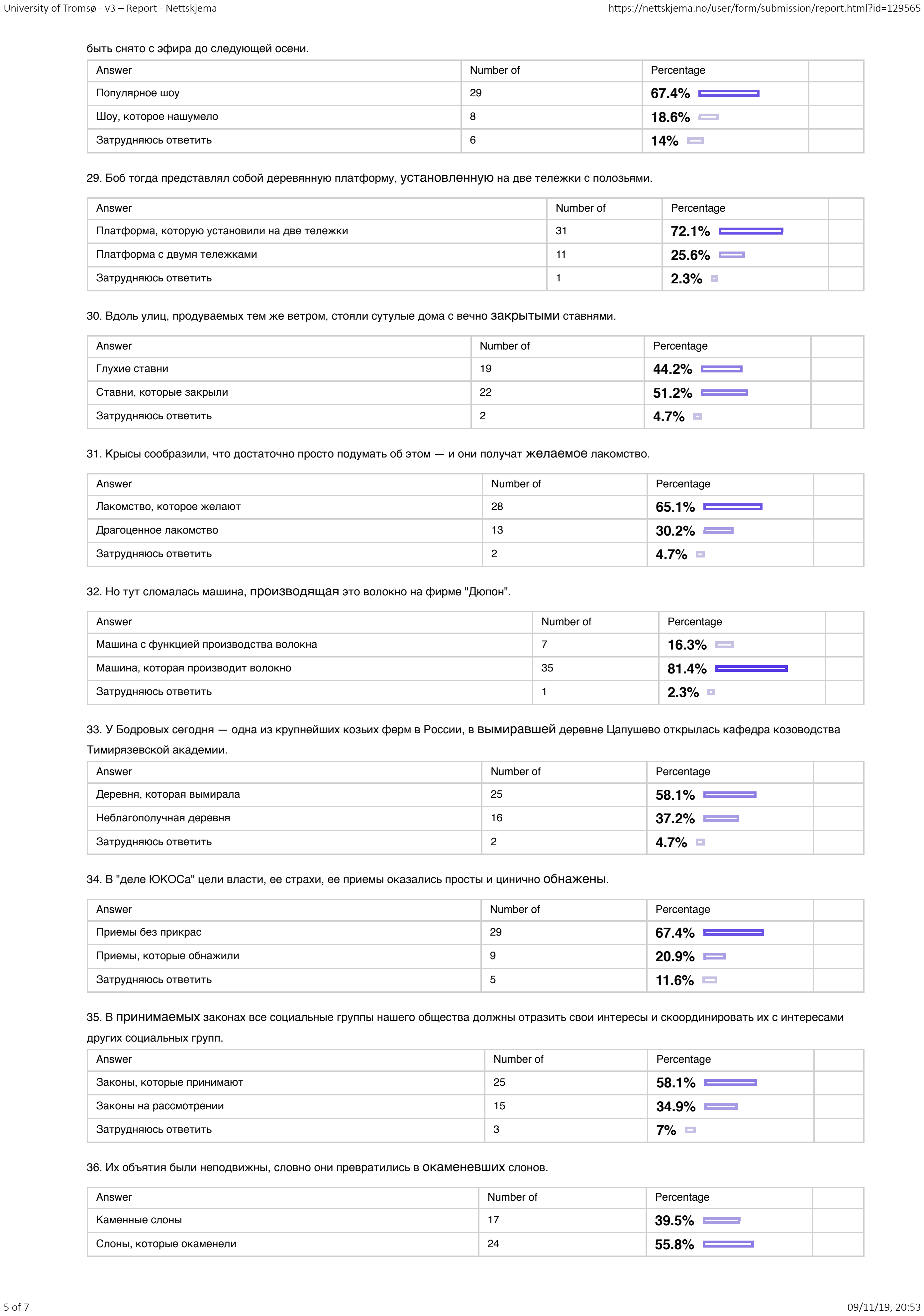
26. В наших делах очень важно, чтобы взаимодействующие люди говорили на одном языке и всё друг про друга понимали.

Answer	Number of	Percentage	
Люди в совместной деятельности	21	48.8%	
Люди, которые взаимодействуют друг с другом	21	48.8%	
Затрудняюсь ответить	1	2.3%	

27. Схема простая — подготовленные люди сдают экзамены за настоящих абитуриентов.

Answer	Number of	Percentage	
Люди, которые подготовились	12	27.9%	
Готовые люди (к экзамену)	25	58.1%	
Затрудняюсь ответить	6	14%	

28. Как сообщил накануне сопредседатель правления ABC Entertainment Television Group Лойд Браун, нашумевшее во всем мире шоу может



быть снято с эфира до следующей осени.

Answer	Number of	Percentage	
Популярное шоу	29	67.4%	
Шоу, которое нашумело	8	18.6%	
Затрудняюсь ответить	6	14%	

29. Боб тогда представлял собой деревянную платформу, установленную на две тележки с полозьями.

Answer	Number of	Percentage	
Платформа, которую установили на две тележки	31	72.1%	
Платформа с двумя тележками	11	25.6%	
Затрудняюсь ответить	1	2.3%	

30. Вдоль улиц, продуваемых тем же ветром, стояли сутулые дома с вечно закрытыми ставнями.

Answer	Number of	Percentage	
Глухие ставни	19	44.2%	
Ставни, которые закрыли	22	51.2%	
Затрудняюсь ответить	2	4.7%	

31. Крысы сообразили, что достаточно просто подумать об этом — и они получат желаемое лакомство.

Answer	Number of	Percentage	
Лакомство, которое желают	28	65.1%	
Драгоценное лакомство	13	30.2%	
Затрудняюсь ответить	2	4.7%	

32. Но тут сломалась машина, производящая это волокно на фирме "Дюпон".

Answer	Number of	Percentage	
Машина с функцией производства волокна	7	16.3%	
Машина, которая производит волокно	35	81.4%	
Затрудняюсь ответить	1	2.3%	

33. У Бодровых сегодня — одна из крупнейших козьих ферм в России, в вымиравшей деревне Цапушево открылась кафедра козоводства Тимирязевской академии.

Answer	Number of	Percentage	
Деревня, которая вымирала	25	58.1%	
Неблагополучная деревня	16	37.2%	
Затрудняюсь ответить	2	4.7%	

34. В "деле ЮКОСа" цели власти, ее страхи, ее приемы оказались просты и цинично обнажены.

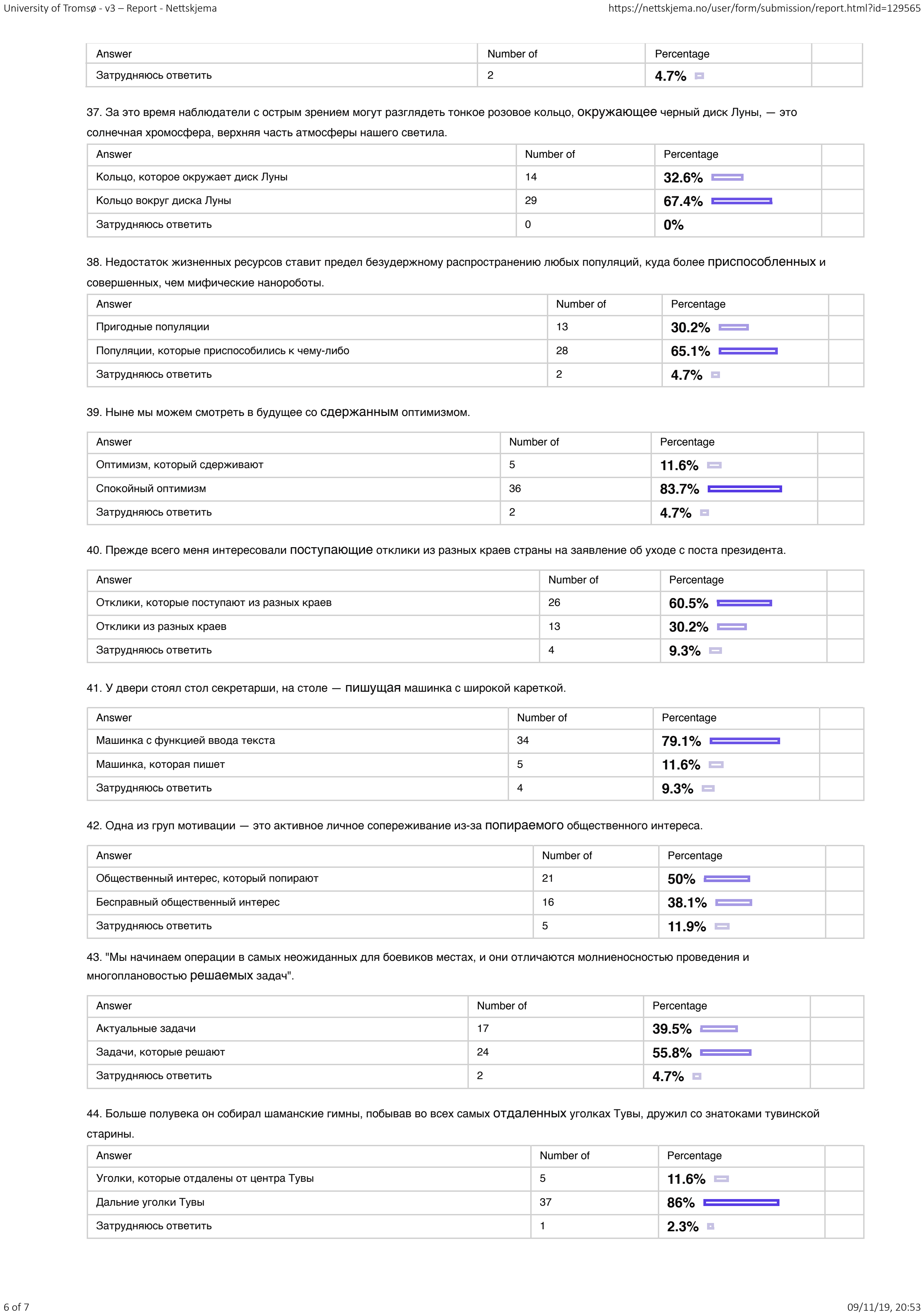
Answer	Number of	Percentage	
Приемы без прикрас	29	67.4%	
Приемы, которые обнажили	9	20.9%	
Затрудняюсь ответить	5	11.6%	

35. В принимаемых законах все социальные группы нашего общества должны отразить свои интересы и скоординировать их с интересами других социальных групп.

Answer	Number of	Percentage	
Законы, которые принимают	25	58.1%	
Законы на рассмотрении	15	34.9%	
Затрудняюсь ответить	3	7%	

36. Их объятия были неподвижны, словно они превратились в окаменевших слонов.

Answer	Number of	Percentage	
Каменные слоны	17	39.5%	
Слоны, которые окаменели	24	55.8%	



Answer	Number of	Percentage	
Затрудняюсь ответить	2	4.7%	

37. За это время наблюдатели с острым зрением могут разглядеть тонкое розовое кольцо, окружающее черный диск Луны, — это солнечная хромосфера, верхняя часть атмосферы нашего светила.

Answer	Number of	Percentage	
Кольцо, которое окружает диск Луны	14	32.6%	
Кольцо вокруг диска Луны	29	67.4%	
Затрудняюсь ответить	0	0%	

38. Недостаток жизненных ресурсов ставит предел безудержному распространению любых популяций, куда более приспособленных и совершенных, чем мифические нанороботы.

Answer	Number of	Percentage	
Пригодные популяции	13	30.2%	
Популяции, которые приспособились к чему-либо	28	65.1%	
Затрудняюсь ответить	2	4.7%	

39. Ныне мы можем смотреть в будущее со сдержанным оптимизмом.

Answer	Number of	Percentage	
Оптимизм, который сдерживают	5	11.6%	
Спокойный оптимизм	36	83.7%	
Затрудняюсь ответить	2	4.7%	

40. Прежде всего меня интересовали поступающие отклики из разных краев страны на заявление об уходе с поста президента.

Answer	Number of	Percentage	
Отклики, которые поступают из разных краев	26	60.5%	
Отклики из разных краев	13	30.2%	
Затрудняюсь ответить	4	9.3%	

41. У двери стоял стол секретарши, на столе — пишущая машинка с широкой кареткой.

Answer	Number of	Percentage	
Машинка с функцией ввода текста	34	79.1%	
Машинка, которая пишет	5	11.6%	
Затрудняюсь ответить	4	9.3%	

42. Одна из груп мотивации — это активное личное сопереживание из-за попираемого общественного интереса.

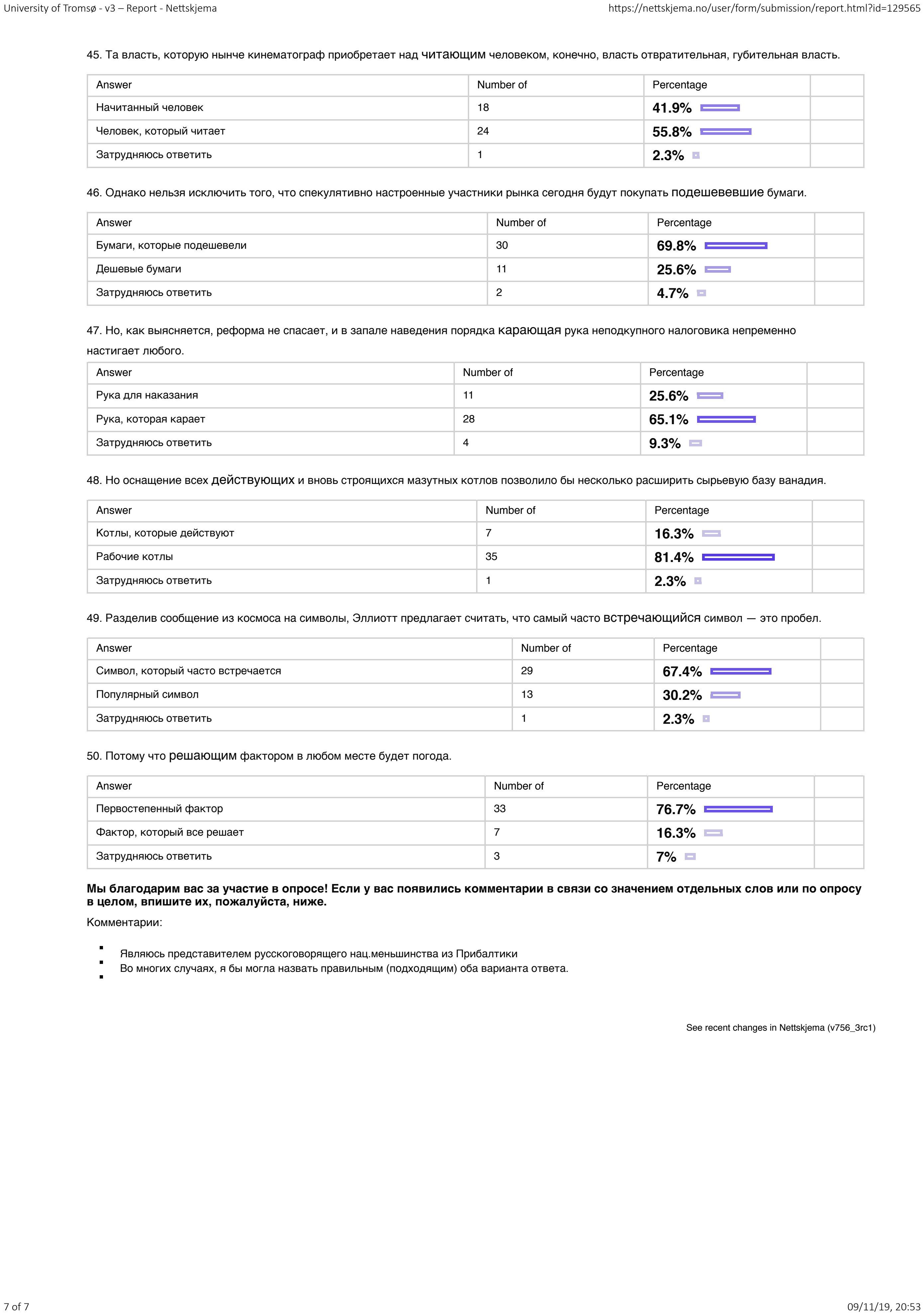
Answer	Number of	Percentage	
Общественный интерес, который попирают	21	50%	
Бесправный общественный интерес	16	38.1%	
Затрудняюсь ответить	5	11.9%	

43. "Мы начинаем операции в самых неожиданных для боевиков местах, и они отличаются молниеносностью проведения и многоплановостью решаемых задач".

Answer	Number of	Percentage	
Актуальные задачи	17	39.5%	
Задачи, которые решают	24	55.8%	
Затрудняюсь ответить	2	4.7%	

44. Больше полувека он собирал шаманские гимны, побывав во всех самых отдаленных уголках Тувы, дружил со знатоками тувинской старины.

Answer	Number of	Percentage	
Уголки, которые отдалены от центра Тувы	5	11.6%	
Дальние уголки Тувы	37	86%	
Затрудняюсь ответить	1	2.3%	



45. Та власть, которую нынче кинематограф приобретает над читающим человеком, конечно, власть отвратительная, губительная власть.

Answer	Number of	Percentage	
Начитанный человек	18	41.9%	
Человек, который читает	24	55.8%	
Затрудняюсь ответить	1	2.3%	

46. Однако нельзя исключить того, что спекулятивно настроенные участники рынка сегодня будут покупать подешевевшие бумаги.

Answer	Number of	Percentage	
Бумаги, которые подешевели	30	69.8%	
Дешевые бумаги	11	25.6%	
Затрудняюсь ответить	2	4.7%	

47. Но, как выясняется, реформа не спасает, и в запале наведения порядка карающая рука неподкупного налоговика непременно настигает любого.

Answer	Number of	Percentage	
Рука для наказания	11	25.6%	
Рука, которая карает	28	65.1%	
Затрудняюсь ответить	4	9.3%	

48. Но оснащение всех действующих и вновь строящихся мазутных котлов позволило бы несколько расширить сырьевую базу ванадия.

Answer	Number of	Percentage	
Котлы, которые действуют	7	16.3%	
Рабочие котлы	35	81.4%	
Затрудняюсь ответить	1	2.3%	

49. Разделив сообщение из космоса на символы, Эллиотт предлагает считать, что самый часто встречающийся символ — это пробел.

Answer	Number of	Percentage	
Символ, который часто встречается	29	67.4%	
Популярный символ	13	30.2%	
Затрудняюсь ответить	1	2.3%	

50. Потому что решающим фактором в любом месте будет погода.

Answer	Number of	Percentage	
Первостепенный фактор	33	76.7%	
Фактор, который все решает	7	16.3%	
Затрудняюсь ответить	3	7%	

Мы благодарим вас за участие в опросе! Если у вас появились комментарии в связи со значением отдельных слов или по опросу в целом, впишите их, пожалуйста, ниже.

Комментарии:

- Являюсь представителем русскоговорящего нац.меньшинства из Прибалтики
- Во многих случаях, я бы могла назвать правильным (подходящим) оба варианта ответа.

See recent changes in Nettskjema (v756_3rc1)

2 Ambiguous sentences

Word form	Sentence	Tag	Answer3 ‘I cannot say’	Answer2 ‘verb’	Answer1 ‘adj’
suščestvujuščix	**sent22**	**ambig**	**30%–70%**	**30%–70%**	**0%–30%**
opozdavšego	sent2	ptcp	70%–100%	0%–30%	0%–30%
ustarevšix	sent8	ptcp	70%–100%	0%–30%	0%–30%
podeševevšie	sent46	ptcp	30%–70%	0%–30%	0%–30%
vstrečajuščijsja	sent49	ptcp	30%–70%	0%–30%	0%–30%
karajuščaja	sent47	ptcp	30%–70%	0%–30%	0%–30%
prisposoblennyx	sent38	ptcp	30%–70%	0%–30%	0%–30%
želaemoe	sent31	ptcp	30%–70%	0%–30%	0%–30%
sedejuščij	sent11	ptcp	30%–70%	30%–70%	0%–30%
postupajuščie	sent40	ptcp	30%–70%	0%–30%	0%–30%
vosxodjaščix	sent14	ptcp	30%–70%	30%–70%	0%–30%
prinimaemyx	sent35	ptcp	30%–70%	30%–70%	0%–30%
vymiravšej	sent33	ptcp	30%–70%	30%–70%	0%–30%
zabrošennyj	sent12	ptcp	30%–70%	30%–70%	0%–30%
čitajuščim	sent45	ptcp	30%–70%	30%–70%	0%–30%
rešaemyx	sent43	ptcp	30%–70%	30%–70%	0%–30%
okamenevšix	sent36	ptcp	30%–70%	30%–70%	0%–30%
zakrytymi	**sent30**	**ambig**	**30%–70%**	**30%–70%**	**0%–30%**
popiraemogo	sent42	ptcp	30%–70%	30%–70%	0%–30%
vzaimodejstvujuščie	**sent26**	**ambig**	**30%–70%**	**30%–70%**	**0%–30%**
monopolizirovannuju	sent4	adj	30%–70%	30%–70%	0%–30%
oxlaždajuščie	sent17	adj	30%–70%	30%–70%	0%–30%
istekšie	sent6	adj	30%–70%	0%–30%	0%–30%
podgotovlennye	sent27	adj	0%–30%	30%–70%	0%–30%
pokalečennyx	sent15	adj	0%–30%	30%–70%	0%–30%
predpolagaemoe	sent7	adj	0%–30%	70%–100%	0%–30%
obnaženy	sent34	ptcp	0%–30%	30%–70%	0%–30%
našumevšee	sent28	adj	0%–30%	30%–70%	0%–30%
perevernutom	sent24	adj	0%–30%	70%–100%	0%–30%
dejstvujuščix	sent48	adj	0%–30%	70%–100%	0%–30%
blagouxajuščix	sent18	adj	0%–30%	70%–100%	0%–30%
prosvečivajuščie	sent5	adj	0%–30%	70%–100%	0%–30%
skanirujuščie	sent5	adj	0%–30%	70%–100%	0%–30%
podyxajuščej	sent1	adj	0%–30%	70%–100%	0%–30%
ostavšeesja	sent16	adj	0%–30%	70%–100%	0%–30%
ošelomljajuščej	sent19	adj	0%–30%	70%–100%	0%–30%
osvedomlennogo	sent20	adj	0%–30%	70%–100%	0%–30%
rasprostranennymi	sent21	adj	0%–30%	70%–100%	0%–30%
rasširennogo	sent25	adj	0%–30%	70%–100%	0%–30%

Table H.1: Classified sentences with the tags *ptcp*, *adj*, and *ambig* assigned to the ambiguous word forms after the completion of the survey.

Appendix I

CG rules for ambiguous participles

Table I.1 presents a list of the CG rules used in the disambiguation experiment in Chapter 5. The rules are numbered according to the order in which they appear in the Russian CG. I provide most of the rules accompanied by a brief description, and optional examples illustrating an ambiguous participle in bold.

#	CG rules
1	REMOVE:PTCP-SpecC-N1 N IF (0 V)(0 N + $$NGDAIP)(-1 Comm LINK -1 N + $$NGDAIP) e.g., *правила, **говорящие** “да"и “нет но уже вознесённого на гребень успеха.*
2	REMOVE:PTCP-SpecC-N2 N IF (0 V) (0 N + $$NGDAIP)(-1 PurposeAdv)(1 N + $$NGDAIP)
3	REMOVE:PTCP-SpecC- N3 N IF (0 V) (0 N + $$NGDAIP)(-1 Comm LINK -1 N + $$NGDAIP)(1 Num LINK 1 N)
4	REMOVE:PTCP-SpecC-N4 V IF (0 V) (0 N + $$NGDAIP)(-1 Comm LINK -1 N)(1 Pr) e.g., *написанные руководителями команд, **участвующих** в RoboCup*
5	REMOVE:PTCP-SpecC-N5 V IF (0 V + $$NGDAIP) (0 N + $$NGDAIP)(-1 Comm LINK -1 N)(1 Pr)
6	REMOVE:PTCP-SpecC-N6.1 V IF (0 V + Gen) (0C N + Gen)(0C A + Gen)(-1 N)(1 V)
7	REMOVE:PTCP-SpecC-N6.2 A IF (0C A + Gen) (0C N + Gen)(0C A + Gen)(-1 N)(1 V) e.g., *подавляющее боль шинство **погибших** составили вооруженные боевики*
8	REMOVE:PTCP-SpecC-N7.1 N IF (0 Ptcp + $$NGDAIP + $$NBR)(0 N + Gen)(-1 N)(1 Pr LINK 1 N LINK 1 N + $$NGDAIP + $$NBR) e.g., *несколько бурых бугров, **выдающихся** из воды*
9	REMOVE:PTCP-SpecC-N7.2 V IF (0 Ptcp + $$NGDAIP) (0 N + Gen)(-1 N)(*1 Pr BARRIER N + $$NGDAIP)
10	REMOVE:PTCP-SpecC-N7.3 A IF (0 A + Gen) (0 N + Gen)(-1 N)(1 Pr) e.g., *интервью **ведущих** с героем выпуска*
11	REMOVE:PTCP-SpecC-N8.1 V IF (0 V + Gen) (0 N + Gen)(-1 N OR Pr)(1 N + Gen) e.g., *интервью **ведущих** школы злословия*, *интервью **ведущих** "школы злословия"*
12	REMOVE:PTCP-SpecC-N8.2 A IF (0 A + Gen) (0 N + Ge n)(-1 N OR Pr)(1 N + Gen)
13	REMOVE:PTCP-SpecC-N8.3 V IF (0 V + Gen) (0 N + Gen)(-1 N OR Pr)(1 Quot LINK 1 N + Gen)
14	REMOVE:PTCP-SpecC-N8. 4 A IF (0 A + Gen) (0 N + Gen)(-1 N OR Pr)(1 Quot LINK 1 N + Gen)
15	REMOVE:PTCP- SpecC-N8.5 V IF (0 V + Gen) (0 N + Gen)(-1 N OR Pr)(1 Quot)
16	REMOVE:PTCP-SpecC-N8.6 A IF (0 A + Gen) (0 N + Gen)(-1 N OR Pr)(1 Quot)
17	REMOVE:PTCP-SpecC-N9 .1 A OR V IF (0 A + Dat) (0 N + Dat)(-1 Pron LINK -1 N LINK -1 V) e.g., *Владимир Путин устроил разнос своим **подчиненным** за скачок цен на нефтепродукты . . .*

Table I.1 – continued on the next page

#	CG rules
18	REMOVE:PTCP-SpecC-N9. 2 A OR V IF (0 A + Dat) (0 N + Dat)(-1 N LINK -1 V)
19	REMOVE:PTCP-SpecC-N9.3 V IF (0 A + Dat) (0 N + Dat)(-1 Pron LINK -1 N LINK -1 V)
20	REMOVE: PTCP -SpecC-N9.4 V IF (0 A + Dat) (0 N + Dat)(-1 N LINK -1 V)
21	REMOVE:PTCP-Spec C-N10 N IF (0 V) (0 N + $$NGDAIP)((1 N + $$NGDAIP) OR (1 Pr LINK 1 N LINK 1 N + $$NGDAIP)) preposition phrase as an adjunct e.g., ***Курящие*** *в стороне люди . . .*
22	SELECT:PTCP-SpecC-N11 N IF (0 V) (0 N)(1 S-BOUNDARY)((-1 Num)OR (-1 Num LI NK -1 V)) select a noun if it is preceded by a numeral optionally preceded by a verb
23	REMOVE:PTCP-SpecC-V1.1 V IF (0 Ptcp)(0 A + $ $NGDAIP)(NOT 0 Adv)(NOT 0 N) (1 N + $$NGDAIP LINK NOT 1 Pr LINK 1 N OR Pron) stand-alone use of an ambig. participle e.g., *подготовка* ***квалифицированных*** *учителей – русистов в Азербайджане*
24	REMOVE:PTCP-SpecC-V1.2 V IF (0 A + $$NGDAIP)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N + $$NGDAIP LINK NOT 1 Pr LINK 1 A LINK 1 N OR Pron) stand-alone use of an ambig. participle e.g., *Вскоре после этого* ***волнующего*** *события было заявлено . . .*
25	REMOVE:PTCP-SpecC-V2.1 V IF (0 A + $$NGDAIP)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv LINK -1 Comm LINK -1 N + $$NGDAIP OR Pron + $$NGDAIP) OR (-1 MeasureDegreeAdv LINK -1 Ne LINK -1 Comm LINK -1 N + $$NGDAIP OR Pron + $$NGDAIP))(1 CC LINK NOT 1 Pron + NOTNOMPREP OR N + NOTNOMPREP) adverb of measure/degree, stand-alone use e.g., *спросил философ, настолько* ***удивленный*** *и обиженный, насколько было это возможно*
26	REMOVE:PTCP-SpecC-V2.2 V IF (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 CC LINK -1 A or V LINK -1 MeasureDegreeAdv)OR(-1 CC LINK -1 A LINK -1 MeasureDegreeAdv LINK -1 Ne))(1 Comm) adverb of measure/degree, stand-alone use e.g., *спросил философ, настолько удивленный и* ***обиженный****, насколько было это возможно при его характере*
27	REMOVE:PTCP-SpecC-V3 V IF (0 A + $$NGDAIP + $$NBR)(0 Ptcp + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv OR Det) OR (-1 Ne LINK -1 MeasureDegreeAdv OR Det) OR (-1 V))(1 Pron + $$NGDAIP + $$NBR OR N + $$NGDAIP + $$NBR) stand-alone use, ambig. participle + ambig. participle + head noun e.g., *как это обычно делают* ***курящие*** *люди в задумчивости*
28	REMOVE:PTCP-SpecC-V4.1 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 Adv/Cmpar) OR (-1 Ne LINK -1 Adv/Cmpar))(NOT 1 N + NOTNOMPREP OR Pron + NOTNOMPREP) adverb of comparative degree e.g., *На репетициях видели гораздо более* ***собранного*** *и звонкоголосого певца . . .*
29	REMOVE:PTCP-SpecC-V4.2 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 Samyj) OR (-1 Ne LINK -1 Samyj))(NOT 1 N + NOTNOMPREP OR Pron + NOTNOMPREP) adverb of superlative degree e.g., *На репетициях видели самого* ***собранного*** *и звонкоголосого певца . . .*
30	REMOVE:PTCP-SpecC-V5 V IF (0 A + $$NGDAIP)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 N + $$NGDAIP OR Pron + $$NGDAIP) OR (-1 Ne LINK -1 N + $$NGDAIP))(NOT 1 Punct)(NOT 1 Pr LINK 1 Loc) e.g., (1) *А любимее всего – финал* ***открытый****.* (2) *Пожилая пара . . . находилась под впечатлением от архитектуры* ***бывшей*** *пивоварни.* (3) *но главное* ***действующее*** *лицо и "деятель" революции отсутствует . . .*

Table I.1 – continued on the next page

#	CG rules
31	REMOVE:PTCP-SpecC-V6 V IF (0 V)(0 A)(0 Ptcp)(NOT 0 Pred)(NOT 0 Adv)(NOT 0 N)((-1 CC) OR (-1 Ne LINK -1 CC))(1 S-BOUNDARY or (WORD S-BOUNDARY)) e.g., (a) *Методика наших вычислений — упрощенная и схематичная, но зато прозрачная и* ***открытая***. (b) *Все грязное и* ***обтрёпанное*** *. . .*
32	REMOVE:PTCP-SpecC-V7.1 V IF (0 A + $$NGDAIP)(0 Ptcp + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv) OR (-1 MeasureDegreeAdv LINK -1 Ne))(1 Pr LINK 1 N + Loc OR Pron + Loc OR Det + Loc OR Ptcp + Loc LINK 1 N + $$NGDAIP) adverb of measure/degree, preposition phrase as an adjunct e.g., *Это был достаточно* ***уверенный*** *в себе человек.*
33	REMOVE:PTCP-SpecC-V7.2 V IF (0 A + $$NGDAIP)(0 Ptcp + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv)OR(-1 MeasureDegreeAdv LINK -1 Ne))(1 Pr LINK 1 Det + Loc LINK 1 N + Loc LINK 1 N + $$NGDAIP) adverb of measure/degree + ambig. participle + preposition phrase + head noun; adverb of measure/degree + preposition +noun/pronoun e.g., *Это был достаточно* ***осведомленный*** *в этом вопросе человек.*
34	REMOVE:PTCP-SpecC-V7.3 V IF (0 A + $$NGDAIP)(0 Ptcp + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv)OR(-1 MeasureDegreeAdv LINK -1 Ne))(1 Pr LINK 1 A + Loc LINK 1 N + Loc LINK 1 N + $$NGDAIP) adv. modifier + ambig. participle + preposition phrase + head noun; preposition + adj + noun + head noun e.g., *Это был достаточно* ***осведомленный*** *в привычном распорядке человек.*
35	REMOVE:PTCP-SpecC-V7.4 V IF (0 A + $$NGDAIP)(0 Ptcp + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv) OR (-1 MeasureDegreeAdv LINK -1 Ne))(1 Pr LINK 1 ModLoc LINK 1 N + Loc LINK 1 N + $$NGDAIP) adverb of measure/degree e.g., *Это был не достаточно* ***уверенный*** *в пяти случаях человек.*
36	REMOVE:PTCP-SpecC-V8 V IF (0 Ptcp + $$NGDAIP)(0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)((-1 CC LINK -1 A OR N) OR (-1 V))(1 N + $$NGDAIP - Ins) ambig. participle is part of verbal phrase, preposed use, no complement e.g., *Поэтому к гидроизоляционным материалам, используемым на подобных кровлях, предъявляются* ***повышенные*** *требования.*
37	REMOVE:PTCP-SpecC-V9 V IF (0 Ptcp + Gen)(0 A + Gen)(NOT 0 Adv)(NOT 0 N)((-1 Adv - TempLocAdv) OR (-1 Adv - TempLocAdv LINK -1 Adv) OR (-1 CC LINK -1 A OR N) OR (-1 N) OR (-1 N LINK -1 Ne))(1 N + Gen) stand-alone use, adverbial modifier e.g., *В начале* ***текущего*** *месяца АвтоВАЗ подписал договор о намерениях с банком «Сосьете Женераль Восток» (BSGV) . . .*
38	REMOVE:PTCP-SpecC-V10 V IF (0 Ptcp) (0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)(-1 CC OR Comm LINK -1 A)(1 Comm OR CC) postposed adjective e.g., *Народ немолодой,* ***почтенный****, в большинстве семейный, а приходится как бы играть в революцию . . .*
39	REMOVE:PTCP-SpecC-V11 V IF (0 Ptcp)(0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)(-1 CC OR Comm LINK -1 A)(1 Comm OR CC) postposed adjective e.g., *Огромные размеры дореволюционных фабричных залов,* ***захватывающий*** *вид на Кремль и Пречистенскую набережную.*

Table I.1 – continued on the next page

#	CG rules
40	REMOVE:PTCP-SpecC-V12 V IF (0 Ptcp)(0 A + Pred)(NOT 0 Adv)(NOT 0 N)(-1 N + $$NGDAIP + $$NBR OR Pron + $$NGDAIP + $$NBR LINK -1 Comm)(1 Comm) predicative use of ambig. participle e.g., *Но задатки виртуоза, я **уверена**, даны вам природой ...*
41	REMOVE:PTCP-SpecC-V13 V IF (0 Ptcp + Ins)(0 A + Ins)(NOT 0 Adv)(NOT 0 N)(-1 Pron + Acc OR N + Acc LINK -1 V)(1 Punct) verbal phrase e.g., *К сожалению , приходится признать , что так же мало понимания и осознанности у многих даже из тех , кто считает себя **верующими** , в том , ...*
42	REMOVE:PTCP-SpecC-V14 V IF (0 A + $$NGDAIP + $$NBR)(0 Ptcp + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Punct OR CS OR CC OR Pr OR Ili)(1C N + $$NGDAIP + $$NBR) stand-alone use e.g., *Зато с каким упорством многие соотечественники нападают на желающих порадовать **любимую** девушку букетом цветов ...*
43	REMOVE:PTCP-SpecC-A1.1 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 Cop)(1 S-BOUNDARY) use of participles as copular verb objects, predicative use of participles e.g., *Машина была **угнана** ...*
44	REMOVE:PTCP-SpecC-A1.2 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 CC OR Comm LINK -1 Ptcp OR A LINK -1 Cop) copular verb objects, predicative use of participles e.g., *Машина была **угнана** и продана ...*
45	REMOVE:PTCP-SpecC-A1.3 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 Cop)(1 Pr) copular verb objects, predicative use of participles e.g., *Кот был **загнан** в угол.*
46	REMOVE:PTCP-SpecC-A1.4 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 CC LINK -1 Cop)(1 Comm LINK 1 CC) use of participles as copular verb objects, predicative use of ambig. participle e.g., *Машина была не только **найдена**, но и возвращена владельцу.*
47	REMOVE:PTCP-SpecC-A1.5 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 CC lINK -1 Comm LINK -1 Ptcp + Pred) use of participles as copular verb objects, predicative use of ambig. participle e.g., *Машина была не только найдена, но и **возвращена** владельцу.*
48	REMOVE:PTCP-SpecC-A1.6 A IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 CC/A LINK -1 Comm LINK -1 Ptcp + Pred LINK -1 Ne LINK -1 Cop) use of participles as copular verb objects, predicative use of ambig. participle e.g., *Двери были не открыты, а **закрыты**.*
49	REMOVE:PTCP-SpecC-A1.7 A IF (0 Ptcp)(0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N) ((-1 Ne LINK -1 Comm LINK -1 N + $$NGDAIP OR Pron + $$NGDAIP) OR (-1 Comm LINK -1 N + $$NGDAIP))((1 N + NOTNOMPREP OR Pron + NOTNOMPREP) OR (1 CC LINK 1 Ne LINK 1 Adv LINK 1 Pr LINK 1 N OR Pron)OR (1 Pr LINK 1 N OR Pron)) adjuncts, complements e.g., *Поведение людей, **окружающих** госпожу Чайковскую, показалось мне довольно подозрительным ...*

Table I.1 – continued on the next page

#	CG rules
50	REMOVE:PTCP-SpecC-A2 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)((-1 Ne LINK -1 Comm LINK *-1 N + $$NGDAIP + $$NBR OR Pron + $$NGDAIP + $$NBR)OR (-1 Comm LINK *-1 N + $$NGDAIP + $$NBR))((1 N + NOTNOMPREP OR Pron + NOTNOMPREP) OR (1 CC LINK 1 Ne LINK 1 Adv LINK 1 Pr LINK 1 N OR Pron)OR (1 Pr LINK 1 N OR Pron)) complement e.g., *Кстати, недавно многие задышали посвободнее — ВАЗ оплатил долги за поставки, **совершенные** в прошлом (!) году.*
51	REMOVE:PTCP-SpecC-A3 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)((-1 N + Acc LINK -1 Adv LINK -1 Comm LINK -1 N + $$NGDAIP + $$NBR)OR (-1 Pron + Acc LINK -1 Adv LINK -1 Comm LINK -1 N + $$NGDAIP + $$NBR)OR (-1 Det + Acc LINK -1 Adv LINK -1 Comm LINK -1 N + $$NGDAIP + $$NBR))(1 Comm) postposed use, adverbs (including measure/degree) e.g., *Отсюда его знаменитые бегства — то, что так поражало людей, мало его **знающих**, и нравилось женщинам . . .*
52	REMOVE:PTCP-SpecC-A4 A IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 N LINK -1 Pr) OR (-1 N LINK -1 Det OR A LINK -1 Pr) OR (-1 N LINK -1 Pr) OR (-1 Adv))(1 Pr LINK 1 N) postposed use, preposition phrase as an adjunct e.g., *и, заприметив мирно **курящих** в стороне мальчишек, подгребла к ним*
53	REMOVE:PTCP-SpecC-A5 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Adv LINK -1 Comm LINK -1 N + $$NGDAIP + $$NBR OR Pron + $$NGDAIP + $$NBR)(1 CC LINK 1 N + NOTNOMPREP OR Pron + NOTNOMPREP) postposition, adverb of measure/degree, instrumental complement e.g., *ответил он, весьма **удивленный** не то нашей осведомленностью, не то нахальством*
54	REMOVE:PTCP-SpecC-A6 A IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 MeasureDegreeAdv) OR (-1 MeasureDegreeAdv LINK -1 Adv) OR (-1 Ne LINK -1 MeasureDegreeAdv) OR (-1 Ne LINK -1 MeasureDegreeAdv LINK -1 Adv))(1 Pron + Ins OR Pron + Poss + Ins OR N + Ins OR A + Ins OR Det + Ins) adverb of measure and degree e.g., *Как правило, функционал подобного рода сайтов настолько **продуман** его разработчиками, что на интуитивном уровне понятен . . .*
55	REMOVE:PTCP-SpecC-A7 A IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 CC) OR (-1 Ne LINK -1 CC) OR (-1 MeasureDegreeAdv) OR (-1 MeasureDegreeAdv LINK -1 Adv) OR (-1 Ne LINK -1 MeasureDegreeAdv) OR (-1 Ne LINK -1 MeasureDegreeAdv LINK -1 Adv))(1 CC LINK 1 A + NOTNOMPREP OR Pron + NOTNOMPREP OR Det + NOTNOMPREP LINK 1 N + NOTNOMPREP) postposed, adverbs of measure/degree e.g., *ответил он, весьма **удивленный** не то нашей осведомленностью, не то нахальством*
56	REMOVE:PTCP-SpecC-A8 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 MeasureDegreeAdv LINK -1 Comm LINK -1 Pron + $$NGDAIP + $$NBR OR N + $$NGDAIP + $$NBR) (1 CC LINK 1 N + Ins OR Pron + Ins) postposed, adverb of measure/degree, instrumental complement e.g., *ответил он, весьма **удивленный** не то нашей осведомленностью, не то нахальством*
57	REMOVE:PTCP-SpecC-A9 A IF (0 Ptcp)(0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)(1 Pr LINK 1 N LINK 1 N + $$NGDAIP) preposed use, preposition phrase as an adjunct e.g., *Длинные, красивые, **блестящие** на солнце волосы - это мечта любой женщины.*

Table I.1 – continued on the next page

#	CG rules
58	REMOVE:PTCP-SpecC-A10 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Comm LINK -1 N + $$NGDAIP + $$NBR OR Pron + $$NGDAIP + $$NBR)(1 QuotPlain LINK 1 WORD) complement e.g., *не выработались окончательные правила, **говорящие** “да” и “нет”, но уже вознесённого на гребень успеха*
59	REMOVE:PTCP-SpecC-A11 A IF (0 Ptcp)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Comm LINK -1 N)(1 Adv OR Kak) postposed position, complement e.g., *Учеников, **мыслящих** иначе, мы готовы немедленно занести в списки нерадивых . . .*
60	REMOVE:PTCP-SpecC-A12 A IF (0 Ptcp)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Pr LINK -1 V)(1 N + Dat LINK 1 N + Loc OR N + Prp) preposed position, complement e.g., *когда большинство членов кибуца не работают на **принадлежащих** кибуцу предприятиях*
61	REMOVE:PTCP-SpecC-A13 A IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)((-1 TempLocAdv LINK -1 Pr) OR (-1 TempLocAdv LINK -1 N LINK -1 Pr))(1 Col OR Dash) temporal adverb, preposition phrase preceding an ambiguous word-form e.g., *новый реактор будет возведен на месте уже **существующего** - на юге Сиднея*
62	REMOVE:PTCP-SpecC-A14 A (0 Ptcp + Pred)(0 A + Pred)(NOT 0 Adv)(NOT 0 N)((-1 N) OR (-1 Ne LINK -1 N) OR (-1 TempLocAdv LINK -1 N)) predicative use of short participial forms e.g., *А любимее всего – финал уже **открыт**; А любимее всего – финал не **открыт**.*
63	REMOVE:PTCP-SpecC-A15 A (0 A + $$NGDAIP + $$NBR)(0 Ptcp + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 Pron + Gen LINK -1 Comm LINK -1 N + $$NGDAIP + $$NBR)(1 Comm) postposed use, agentive complement e.g., *Рискну предположить, что люди, его **лишенные**, как раз наиболее отзывчивы на указания.*
64	REMOVE:WPTCP-GenC-Prep A IF (0 A + (<W=MIN>))(0 Ptcp + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 Pr) preposition phrase as an adjunct
65	REMOVE:PTCP-GenC-V1 V IF (0 A + $$NGDAIP + $$NBR)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(-1 N + $$NGDAIP + $$NBR)(1 Comm) postposed position e.g., *человека **говорящего**, но*
66	REMOVE:PTCP-GenC-A1 A IF (0 Ptcp + $$NGDAIP + $$NBR)(0 A + $$NGDAIP + $$NBR)(NOT 0 Adv)(NOT 0 N)(-1 TempLocAdv)(1 N + $$NGDAIP + $$NBR) adverbs of time/place e.g., *заранее **заданными** свойствами*
67	REMOVE:PTCP-GenC-A2 A IF (0 Ptcp + Pred)(0 A + Pred)(NOT 0 Adv)(NOT 0 N)(1 Nom) predicative use, short form e.g., *Вместо новой книги о семье Андрея Синявского вышел фильм, в основу которого **положены** те самые письма писателя . . .*
68	REMOVE:PTCP-GenC-A3 A IF (0 Ptcp)(0 A + $$NGDAIP)(NOT 0 Adv)(NOT 0 N)(-1 TempLocAdv)(1 N + $$NGDAIP) temporal adverb e.g., *Робот, которого доставили вчера, был с заранее **заданными** свойствами.*
69	SELECT:WPTCP-GenC-A4 A + $$NGDAIP IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(-1C Samyj + $$NGDAIP OR Adv/Cmpar OR Naibolee)(0 A OR V)(0 A + (<W=MIN>)) safe lexicalized operations, combination with adverbs of measure and degree

Table I.1 – continued on the next page

#	CG rules
70	SELECT:PTCP-GenC-A5.1 A IF (0 Ptcp)(0 A + $$NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)(-1 A + $$NGDAIP/NBR) synonymic sequence: an adjective followed by an ambig. participle e.g., *Это был хороший* ***знакомый*** *приятель.*
71	SELECT:PTCP-GenC-A5.2 A IF (0 Ptcp)(0 A + $$NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)(1 A + $$NGDAIP/NBR) synonymic sequence: an adjective followed by an ambig. participle e.g., *подтверждавшие порождённую им истину, всхлипывал и вытирал* ***плачущие*** *счастливые глаза*
72	SELECT:PTCP-GenC-A5.3 A IF (0 Ptcp)(0 A + $$NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)(1 Comm LINK 1 A + $$NGDAIP/NBR) synonymic sequence: an ambig. participle followed by an adjective agreeing in the same case with the participle e.g., *В каждом взгляде,* ***испуганном****, ненавидящем, скрытном, таилась смерть.*
73	SELECT:PTCP-GenC-A5.4 A IF (0 Ptcp + $$NGDAIP/NBR)(0 A + $$NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)(1 N + $$NGDAIP + $$NBR LINK 1 Sent) stand-alone use e.g., *Это был* ***образованный*** *человек.*
74	SELECT:PTCP-GenC-A6.1 A IF (0 A + $$NGDAIP/NBR) (0 Ptcp + $$NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)(-1 Punct OR CS OR CC OR Pr)((1 N + $$NGDAIP/NBR) OR (1 N + $$NGDAIP/NBR LINK 1 A + Ins LINK 1 N + Ins)) ambig. participle is part of a collocation e.g., *Имеются планы и задел на* ***следующий*** *год.*
75	SELECT:PTCP-GenC-A6.2 A IF (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(-1 Pr)(1 N + $$NGDAIP/NBR LINK 1 Gen) participial phrase is part of prespositional phrase e.g., *Но что поразило меня более всего, так это сходство молодой женщины с* ***вдовствующей*** *императрицей.*
76	SELECT:PTCP-GenC-N1 N IF (0 N)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(-1 Pr)(1 N + Gen) noun complement e.g., *Согласно* ***данным*** *журнала «Экономист».*
77	SELECT:PTCP-GenC-V1.1 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(1 Pron + Ref) e.g., *ambig. ptcp is used with reflexive pronouns*
78	SELECT:PTCP-GenC-V2 V IF (0 Ptcp)(0 A)(0C NOINS)(NOT 0 Adv)(NOT 0 N)(1 Ins) agentive (instrumental) complement e.g., *«— спрашивает Дмитриев, все более и более удивленный и* ***испуганный*** *словами солдата*
79	SELECT:PTCP-GenC-V3 V IF (0 Ptcp-noPass + $$NOTACC)(0 A + $$NOTACC)(NOT 0 Adv)(NOT 0 N)(1C A + Acc OR N + Acc) complement e.g., *Наряду с авторами популярных интернет-дневников и малоизвестными, но* ***вызывающими*** *явную симпатию ведущими персонажами . . .*
80	SELECT:WPTCP-GenC-V4 V IF (0 Ptcp + (<W=MAX>))(0 A + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(1 Dat) stand-alone use, preposed use e.g., *а фоне заявлений президента о необходимости помощи* ***голодающим*** *детям мира*
81	SELECT:PTCP-GenC-V5 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 Comm)(1 Nom OR Acc) postposed use e.g., *Решение принято вслед за аналогичным решением о цензуре,* ***принятой*** *два месяца назад . . .*

Table I.1 – continued on the next page

#	CG rules
82	SELECT:PTCP-GenC-V6 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(1 Pr LINK 1 N) prepostion phrase as an adjunct e.g., *Особую склонность избегать самостоятельных решений проявляют в японском деловом мире люди, только что **повышенные** в ранге.*
83	SELECT:PTCP-GenC-V7 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(1 Pr) posposed use, complement e.g., *Скандал, **связанный** с эпистолярным трехтомником, — скорее издательский.*
84	SELECT:PTCP-GenC-V8 V IF (0 Ptcp + Pred)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 Cop) predicative use e.g., *однако детально данное учение было **развито** автором в магистерской диссертации*
85	SELECT:PTCP-GenC-V9 V IF (0 Ptcp)(0 A)(NOT 0 Adv)(NOT 0 N)(-1 TempLocAdv) predicative use, temporal adverb e.g., *Мы не сомневались в последствиях и полагали нового товарища уже **убитым**.*
86	SELECT:PTCP-GenC-A1 A + $$NGDAIP IF (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1C* N + $$NGDAIP BARRIER Punct OR Pr OR Lparen OR NGDAIP - $$NGDAIP LINK -1C* A + $$NGDAIP BARRIER Punct OR Pr OR Lparen OR NGDAIP - $$NGDAIP) ambig. participle precedes an adjective e.g., *При сохранении без изменений **существующей** социально-политической системы . . .*
87	SELECT:PTCP-GenC-A2 A IF (0 A)(0 Ptcp)(NOT 0 Pred)(NOT 0 Adv)(NOT 0 N)(-1 MeasureDegreeAdv) no predicative use, adverbs of measure and degree e.g., *Может быть, та подозрительная, вечно **настороженная** напряженность человека . . .*
88	REMOVE:WPTCP-V1.1 V IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Pred)(NOT 0 Adv)(NOT 0 N)(-1 MeasureDegreeAdv)(1 N OR Pron LINK 1 N) adverb of measure and degree e.g., *в качестве специального «бонуса» покупатели получат крайне **озлобленных** «аборигенов»*
89	SELECT:WPTCP-A1.2 A IF (0 A + (<W=MAX>))(0 Ptcp + (<W=MIN>))(NOT 0 Pred)(NOT 0 Adv)(NOT 0 N)(-1 MeasureDegreeAdv) adverbs of measure and degree
90	SELECT:WPTCP-A1.3 A IF (0 A + (<W=MAX>))(0 Ptcp + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(-1 Adv/Cmpar) adverb of measure and degree
91	SELECT:WPTCP-A1.4 A IF (0 A + (<W=MAX>))(0 Ptcp + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(-1 Samyj OR Naibolee) adverb of superlative degree
92	SELECT:WPTCP-A2.1 A IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 V) OR (-1 Pron OR Det LINK -1 V))(1 N OR Pron) stand-alone use, ambig. participle as an object of verbal phrase e.g., *потому и является самым **любимым** праздником*
93	SELECT:WPTCP-A2.2 A IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 Pr) OR (-1 Det OR Pron LINK -1 Pr))(1 N OR Pron) stand-alone use, ambig. participle as an object of preposition phrase e.g., *мы получили в **уходящем** году «пророссийского» президента Украины*
94	SELECT:WPTCP-A2.3 A IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(-1 N)(1 N + Gen) stand-alone use, ambig. participle used as an object of noun phrase e.g., *учет **повышенной** ставки*

Table I.1 – continued on the next page

#	CG rules
95	SELECT:WPTCP-A2.4 A IF (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)((-1 N) OR (-1 Det OR Pron LINK -1 N))(1 N OR Pron) (NOT 1 Ins) stand-alone use of an ambig. participle e.g., *и ещё на многих других* ***бывших*** *промзонах*
96	SELECT:WPTCP-A2.5 A IF (0 A + (<W=MAX>))(0 Ptcp)(0 NGDAIP/NBR)(NOT 0 Adv)(NOT 0 N)((-1 N) OR (-1 Det OR Pron LINK -1 N))(1 N + $$NGDAIP/NBR OR Pron + $$NGDAIP/NBR) stand-alone use of an ambig. participle e.g., *если в английском варианте* ***горящий*** *человек наклонялся вперед в рукопожатии*
97	SELECT:WPTCP-A2.6 A (0 A + (<W=MAX>))(0 Ptcp)(0 BOS)(NOT 0 Adv)(NOT 0 N)(1 N) stand-alone use, an ambig. participle is used at the beginning of sentence e.g., ***Начинающие*** *планеристы обязаны находиться в пределах границ*
98	SELECT:WPTCP-A5.1 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 CC LINK -1 A LINK -1 Comm LINK -1 A)OR(-1 CC LINK -1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
99	SELECT:WPTCP-A5.2 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 Comm LINK -1 A LINK -1 Comm LINK -1 A) OR (-1 Comm LINK -1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
100	SELECT:WPTCP-A5.3 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A LINK -1 Comm LINK -1 A) OR (-1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
101	SELECT:WPTCP-A5.4 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 CC LINK -1 A) OR (-1 CC LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
102	SELECT:WPTCP-A5.5 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A) OR (-1 A LINK -1 Ne)) e.g., *ambig. participle is part of sequence of adjectives*
103	SELECT:WPTCP-A5.6 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A LINK -1 A)) e.g., *ambig. participle is part of a sequence of adjectives*
104	SELECT:WPTCP-A5.7 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 S-BOUNDARY)((-1 Comm LINK -1 A) OR (-1 CC LINK -1 A)) e.g., *ambig. ptcp is part of a sequence of adjectives*
105	SELECT:PTCP-A6.1 A (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 CC LINK -1 A LINK -1 Comm LINK -1 A) OR (-1 CC LINK -1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
106	SELECT:PTCP-A6.2 A (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 Comm LINK -1 A LINK -1 Comm LINK -1 A) OR (-1 Comm LINK -1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
107	SELECT:PTCP-A6.3 A (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A LINK -1 Comm LINK -1 A) OR (-1 A LINK -1 Comm LINK -1 A LINK -1 Ne)) e.g., *ambig. participle is part of a sequence of adjectives*
108	SELECT:PTCP-A6.4 A (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 CC LINK -1 A) OR (-1 CC LINK -1 A LINK -1 Ne)) ambig. participle is part of a sequence of adjectives e.g., *и* ***согласованной*** *стратегии выхода*

Table I.1 – continued on the next page

#	CG rules
109	SELECT:PTCP-A6.5 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A) OR (-1 A LINK -1 Ne)) -s- e.g., *ambig. participle is part of a sequence of adjectives*
110	SELECT:PTCP-A6.6 A (0 A + (<W=MAX>))(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 N OR Pron)((-1 A LINK -1 A)) -s- e.g., *ambig. participle is part of a sequence of adjectives*
111	SELECT:PTCP-A6.7 A (0 A)(0 Ptcp)(NOT 0 Adv)(NOT 0 N)(1 S-BOUNDARY)((-1 Comm LINK -1 A) OR (-1 CC LINK -1 A)) e.g., *ambig. participle is part of sequence of adjectives*
112	REMOVE:WPTCP-A-Pred A IF (0 A)(0 Ptcp + Pred + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) predicative use
113	REMOVE:PTCP-A-Pred A IF (0 A)(0 Ptcp + Pred)(NOT 0 Adv)(NOT 0 N) predicative use
114	REMOVE:WPTCP-Ne-V A IF (0 A)(0 Ptcp + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(-1 Ne) negation particle
115	REMOVE:WPTCP-V4.1 V IF (0 A)(0 Ptcp + PrsAct + IV + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)((NOT 1 S-BOUNDARY) OR (NOT 1 N + Dat) OR (NOT 0 Pr - Pr/Iz)) active present adjectivized participles
116	REMOVE:WPTCP-V4.2 V IF (0 A)(0 Ptcp + PrsAct + Impf + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)((NOT 1 S-BOUNDARY) OR (NOT 1 N + Dat) OR (NOT 0 Pr - Pr/Iz)) active present ambig. participles e.g., ***скучающий***
117	REMOVE:WPTCP-V5.1 V IF (0 A)(0 Ptcp + PstAct + V/PstActPerf + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) past active ambig. participles, suffixes -vš-, -š-
118	REMOVE:WPTCP-V5.2 V IF (0 A)(0 Ptcp + PstAct + Perf + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) past active ambig. participles
119	REMOVE:WPTCP-V5.3 V IF (0 A)(0 Ptcp + PstAct + Perf + IV + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) past active ambig. participles
120	REMOVE:WPTCP-V6.1 V IF (0 A)(0 Ptcp + PstPss + Perf + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(1 N + NOINS OR Pron + NOINS) passive past ambig. participles with complement e.g., ***уважаемый, воображаемый***
121	REMOVE:WPTCP-V6.2 V IF (0 A)(NOT 0 Adv)(NOT 0 N)(0 Ptcp + PrsPss + Perf + (<W=MIN>)) present passive ambig. participles without complement e.g., ***уважаемый, воображаемый***
122	REMOVE:WPTCP-V7.2 V IF (0 A)(0 Ptcp + PstPss + Perf - Pred + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) passive past ambig. participles without complement (non-predicative use), perfective use
123	REMOVE:WPTCP-V7.3 V IF (0 A)(0 Ptcp + PstPss + Perf + V/PstPssPerf - Pred + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) passive past ambig. participles without complement (non-predicative use), use with suffixes *-nn/t-*
124	REMOVE:WPTCP-V7.1 V IF (0 A)(0 Ptcp + PstPss - Pred + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) past passive ambig. participles, stand-alone use
125	REMOVE:WPTCP-A2.1 A IF (0 A)(0 PresActv + V/Ref + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) active present ambig. participles, with suffix *-sja-*
126	REMOVE:WPTCP-A2.2 A IF (0 A)(0 PresActv + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) present active participles, transitive use

Table I.1 – continued on the next page

#	CG rules
127	REMOVE:WPTCP-A3.1 A IF (0 A)(0 Ptcp + PastActv + V/PstActPerf-Nu + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) past active participles, suffix *-nu-*
128	REMOVE:WPTCP-A3.2 A IF (0 A)(0 Ptcp + PstAct + Impf + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) past active participles, transitive imperfective use
129	REMOVE:WPTCP-A4.1 A IF (0 A)(0 Ptcp + PrsPss + Impf + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles with complement, imperfective transitive
130	REMOVE:WPTCP-A4.2 A IF (0 A)(0 Ptcp + PrsPss + Impf + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles with complement, imperfective use
131	REMOVE:WPTCP-A4.3 A IF (0 A)(0 Ptcp + PrsPss + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles with complement, transitive use
132	REMOVE:WPTCP-A4.4 A IF (0 A)(0 Ptcp + PrsPss + V/PssPrs + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles with complement, used with the suffix *-myj-*
133	REMOVE:WPTCP-A4.5 A IF (0 A)(0 Ptcp + PrsPss + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles with complement
134	REMOVE:WPTCP-A4.6 A IF (0 A)(0 Ptcp + PrsPss + Impf + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) present passive adjectivized participles without complement, imperfective transitive use
135	REMOVE:WPTCP-A4.7 A IF (0 A)(0 Ptcp + PrsPss + Impf + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) present passive adjectivized participles without complement, imperfective use
136	REMOVE:WPTCP-A4.8 A IF (0 A)(0 Ptcp + PrsPss + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) present passive adjectivized participles with complement, transitive use
137	REMOVE:WPTCP-A4.9 A IF (0 A)(0 Ptcp + PrsPss + V/PssPrs + (<W=MAX>))(NOT 0 Adv)(NOT 0 N)(1 N + Ins OR Pron + Ins) present passive participles without complement, used with the suffix *-myj-*
138	REMOVE:WPTCP-V8 V IF (0 A)(0 Ptcp + Pass + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(NOT 0 Pred)(NOT 1 S-BOUNDARY) passive voice
139	REMOVE:WPTCP-V9 V IF (0 A)(0 PrcPres + (<W=MIN>))(NOT 0 Adv)(NOT 0 N) present tense
140	REMOVE:WPTCP-V10 V IF (0 A)(0 Ptcp + Perf + (<W=MIN>))(NOT 0 Adv)(NOT 0 N)(NOT 0 Pred) perfective aspect
141	REMOVE:WPTCP-A5 A IF (0 A)(0 PrcPast + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) past tense
142	REMOVE:WPTCP-A6 A IF (0 A)(0 Ptcp + Impf + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) imperfective aspect
143	REMOVE:WPTCP-A7 A IF (0 A)(0 Ptcp + TV + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) transitive use
144	REMOVE:WPTCP-A8 A IF (0 A)(0 Ptcp + Actv + (<W=MAX>))(NOT 0 Adv)(NOT 0 N) active voice
145	SELECT:maxweight (<W=MAX>) IF (0 A)(0 Ptcp) finite weighted rule: select the reading with the highest weight

Table I.1: Rules for resolving the ambiguity of participles available in the Russian CG. The rules are presented in the order in which they appear in the Russian CG.