Short-term morphosyntactic change: the development of the Russian predicate adjective 1800 – 2000

by

Alison Jane Long

Submitted for the degree of Doctor of Philosophy

Faculty of Arts and Human Sciences
University of Surrey
March 2010
© Alison Jane Long 2010
This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.

Alison Long
University of Surrey
March 2010
Abstract

It is well established that all languages undergo change. Such changes have been seen in the Russian language in a relatively short period of time, particularly in the area of morphosyntax. In order to quantify and examine such changes data from the Russian language was used (the Barentsen Corpus). The aim of his thesis is to quantify change in Russian predicate adjective usage over a two hundred year period, 1801 – 2000, and improve upon the statistical analyses of corpora. A Russian predicate adjective may take one of three forms; the short form (SF), the long form nominative (LF-N) and the long form instrumental (LF-I). It is widely recognised that short form usage is declining in Russian. This study set out to determine how and where this decline was taking place over the two hundred year period. This was done through examination of examples taken from the Barentsen Corpus, which were then divided according to pre-decided criteria and analysed statistically. The results were divided according to the specific grammatical element examined; verbs, nouns and adjectives. Overall, it was found, as expected, that the short form is in decline. Statistical analysis enabled the nature and time of the change, including its statistical significance, to be identified within the time period examined. Through the statistical analysis used in this study it was shown that specific areas of change were identified. For example, the increasing occurrence of the LF-I with semi-copula verbs was observed, and that different rates of change were observed in the time period examined depending on grammatical features and some lexical items (a group of adjectives favouring the SF). Therefore, the methods used in this study allowed the quantification of changes in choice of predicate adjective form, building upon previous work by Benson, Gustavsson and Baženova.
CONTENTS

Declaration of Originality ii
Abstract iii
Table of Contents iv
List of Tables vii
List of Figures viii
Acknowledgements xvi

1 INTRODUCTION 1
1.1 Overview and Context 2
  1.1.1 Overview of the thesis 2
1.2 Historical Linguistics 5
1.3 Predicates in Russian 6
  1.3.1 The influence of literary style 10
1.4 Previous work 16
  1.4.1 Usage 16
  1.4.2 Historical and corpus studies 18
  1.4.3 Semantics 25
  1.4.4 Syntax and agreement 28
  1.4.5 Summary of previous research 34

2 METHODOLOGY 36
2.1 Introduction 37
  2.1.1 Issues in Corpus design 37
2.2 Historical and stylistic investigations 40
  2.2.1 Stylistic and historical studies 40
  2.2.2 The historical evolution of written and speech-based registers 43
  2.2.3 The Barentsen Corpus 45
Analysis of adjective class and predicate form 169
5.4.2 Qualitative analysis of the Corpus: Predicate adjectives with copula and semi-copula verbs 179
5.4.3 Initial Search of the Corpus – Analysis of Collated Adjectives 199
5.4.4 Analysis of change for individual adjectives 203
5.4.5 Stress and stress alternation in the SF 218
5.5 Discussion 222

6 CHAPTER 6 – Conclusions 232

Appendix 246
1801-1850 247
1851-1900 251
1901-1950 255
1951-2000 263

Bibliography 272
LIST OF TABLES

Table 1.1. Examples of the predicate position in Russian. 6
Table 1.2. Comparison of the use of long and short form in prehistoric Slavic and modern Russian (modified from Larsen, 2002). 9
Table 1.3. A model of narrative structure, adapted from Simpson (2004, p.20). 10
Table 1.4. Table of usage of SF adjectives in non-fiction texts. 12
Table 1.5. Table of SF usage in drama texts. 15
Table 2.1. A breakdown of the number of texts contained within particular time periods in the Bartensen Corpus. 47
Table 3.1. Statistical comparison, by χ-squared test, for predicate adjectives with copula and semi-copula verbs. 79
Table 3.2. Results from corpus searches for types that were returned in insufficient numbers for statistical analysis to be carried out. 80
Table 3.3. Statistical comparison, by χ-squared test, of pairs of verbs, in each of the time periods, and the statistical significance of trends. 91
Table 4.1. Statistical comparison, by χ-squared test, for nouns: comparison of frequency of occurrence across each time period, and the statistical significance of trends. 135
Table 4.2. Results from corpus searches for types that were returned in insufficient numbers for statistical analysis to be carried out. 139
Table 4.3. Statistical comparison, by χ-squared test, of pairs of verbs, in each of the time periods, and the statistical significance of trends. 141
Table 5.1. Example of the tables compiled for adjective classes. 170
Table 5.2. Summary of occurrence of SF for each verb in the Corpus. 181
Table 5.3. Summary of occurrence of LF-N for each verb in the Corpus. 182
Table 5.4. Summary of occurrence of LF-I for each verb in the Corpus. 183
Table 5.5. Summary of regression coefficients for different types of line-fitting. 187
Table 5.6. Summary of statistical tests for predicate adjectives where the complete dataset was analysed by collating all adjectives together. 199
Table 5.7. Summary of predicate adjective occurrence in the Corpus. 202
Table 5.8. Common Group Adjective stress. 221
LIST OF FIGURES

Figure 2.1. Examples returned in a search for "SF adjectives with feminine animate nouns" 48

Figure 2.2 An example returned for the search "SF adjectives with inanimate neuter nouns" 48

Figure 2.3. Scatterplot of short forms before and after removal of 12 adjectives. 55

Figure 2.4. A fitted line plot showing the decline of the SF over the time periods examined in this study. 56

Figure 3.1. Change in occurrence with делаться детать'sja (124 total examples; 15, 36, 16 and 87 in each time period). 72

Figure 3.2. Percentage change in occurrence with делаться детать'sja (124 total examples; 15, 36, 16 and 87 in each time period). 73

Figure 3.3. Comparison of the change in occurrence, as numerical (total occurrences) and percentage values, for occurrences of SF делатьться returned from a search of the corpus. 74

Figure 3.4. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs. 82

Figure 3.5. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula verbs. 82

Figure 3.6. Frequency of distribution of SF, LF-N and LF-I across the four time periods for non-reflexive verbs. 83

Figure 3.7. Frequency of distribution of SF, LF-N and LF-I across the four time periods for reflexive verbs. 84

Figure 3.8. Frequency of distribution of SF, LF-N and LF-I across the four time periods for past tense copula verbs. 85

Figure 3.9. Frequency of distribution of SF, LF-N and LF-I across the four time periods for past tense semi copula verbs. 85

Figure 3.10. Frequency of distribution of SF, LF-N and LF-I across the four time periods for present tense semi copula verbs. 87

Figure 3.11. Frequency of distribution of SF, LF-N and LF-I across the four time periods for present tense zero copula verbs. 87

Figure 3.12. Frequency of distribution of SF, LF-N and LF-I across the four time periods for perfective verbs. 88

Figure 3.13. Frequency of distribution of SF, LF-N and LF-I across the four time periods for imperfective verbs. 88
Figure 3.14. Frequency of distribution of SF, LF-N and LF-I across the four time periods for infinitive copula verbs.

Figure 3.15. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1801 - 1850.

Figure 3.16. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1851 - 1900.

Figure 3.17. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1901 - 1950.

Figure 3.18. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1951 - 2000.

Figure 3.19. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1801 - 1850.

Figure 3.20. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1851 - 1900.

Figure 3.21. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1901 - 1950.

Figure 3.22. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1951 - 2000.

Figure 3.23. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1801 - 1850.

Figure 3.24. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1851 - 1900.

Figure 3.25. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1901 - 1950.

Figure 3.26. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1951 - 2000.

Figure 3.27. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense copula vs. present tense semi copula verbs in the time period 1801 - 1850.

Figure 3.28. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense zero copula vs. present tense semi copula verbs in the time period 1851 - 1900.

Figure 3.29. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense copula vs. present tense semi copula verbs in the time period 1901 - 1950.

Figure 3.30. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense copula vs. present tense semi copula verbs in the time period 1951 - 2000.

Figure 3.31. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1801 - 1850.
Figure 3.32. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1851 - 1900.

Figure 3.33. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1901 - 1950.

Figure 3.34. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1951 - 2000.

Figure 3.35. Cross tabulation output and results of $\chi^2$ tests from SPSS® (v.14) for copula and semi copula verbs.

Figure 3.36. Annotated output from the Symmetric Measures test, using SPSS® (v.14). The three boxes highlight the key issues with interpreting this data.

Figure 3.37. Annotated output from the Directional Measures test, using SPSS® (v.14).

Figure 3.38. Crosstabulation output from SPSS® (v.14) for analysis of LF-I and LF-N.

Figure 3.39. Output from the directional (top) and symmetric (bottom) measures analysis of the LF-I and LF-N data for copula and semi copula verbs.

Figure 3.40. Directional (top) and symmetric (bottom) measures for a comparison of SF versus combined LF.

Figure 4.1. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; concrete).

Figure 4.2. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; abstract).

Figure 4.3. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; animate).

Figure 4.4. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; inanimate).

Figure 4.5. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; concrete inanimate).

Figure 4.6. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; concrete animate).

Figure 4.7. Frequency of distribution of SF, LF-N and LF-I across the four time periods for *by itself* (concrete).

Figure 4.8. Frequency of distribution of SF, LF-N and LF-I across the four time periods for *by itself* (abstract).

Figure 4.9. Frequency of distribution of SF, LF-N and LF-I across the four time periods for *by itself* (animate).

Figure 4.10. Frequency of distribution of SF, LF-N and LF-I across the four time periods for *by itself* (inanimate).
Figure 4.11. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (concrete).

Figure 4.12. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (abstract).

Figure 4.13. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (animate).

Figure 4.14. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (inanimate).

Figure 4.15. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1801 - 1850.

Figure 4.16. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1851 - 1900.

Figure 4.17. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1901 - 1950.

Figure 4.18. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1951 - 2000.

Figure 4.19. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1801 - 1850.

Figure 4.20. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1851 - 1900.

Figure 4.21. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1901 - 1950.

Figure 4.22. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1951 - 2000.

Figure 4.23. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1801 - 1850.

Figure 4.24. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1851 - 1900.

Figure 4.25. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1901 - 1950.

Figure 4.26. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1951 - 2000.

Figure 4.27. Comparison of percentage occurrence of SF, LF-N and LF-I for byt’ (concrete vs. abstract) nouns in the time period 1801 - 1850.

Figure 4.28. Comparison of percentage occurrence of SF, LF-N and LF-I for byt’ (concrete vs. abstract) nouns in the time period 1851 - 1900.
Figure 4.29. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (concrete vs. abstract) nouns in the time period 1901 - 1950. 150

Figure 4.30. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (concrete vs. abstract) nouns in the time period 1951 - 2000. 151

Figure 4.31. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (animate vs. abstract) nouns in the time period 1801 - 1850. 151

Figure 4.32. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (animate vs. abstract) nouns in the time period 1951 - 2000. 152

Figure 4.33. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (animate vs. abstract) nouns in the time period 1951 - 2000. 152

Figure 4.34. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (animate vs. abstract) nouns in the time period 1951 - 2000. 153

Figure 4.35. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (concrete vs. abstract) in the time period 1801 - 1850. 156

Figure 4.36. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (concrete vs. abstract) in the time period 1851 - 1900. 156

Figure 4.37. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (concrete vs. abstract) in the time period 1901 - 1950. 156

Figure 4.38. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (concrete vs. abstract) in the time period 1951 - 2000. 157

Figure 4.39. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (animate vs. inanimate) in the time period 1801 - 1850. 158

Figure 4.40. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (animate vs. inanimate) in the time period 1851 - 1900. 158

Figure 4.41. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (animate vs. inanimate) in the time period 1901 - 1950. 159

Figure 4.42. Comparison of percentage occurrence of SF, LF-N and LF-I for semi copula nouns (animate vs. inanimate) in the time period 1951 - 2000. 159

Figure 4.43. Schematic representation of proposed types of language change. 165

Figure 5.1. Comparison of predicate adjectives from 1801 - 2000 in drama texts. 171

Figure 5.2. Relative proportions (%) of predicate adjectives according to class for Griboedov, representing the block 1801 - 1850. 172

Figure 5.3. Relative proportion (%) of lexical items according to adjective class for Griboedov, representing the block 1801 - 1850. 172

Figure 5.4. Relative proportions (%) of predicate adjectives according to class for Turgenev, representing the block 1851 - 1900. 173
Figure 5.5. Relative proportion (%) of lexical items according to adjective class for Turgenev, representing the block 1851 – 1900.

Figure 5.6. Relative proportions (%) of predicate adjectives according to class for Gor’kij, representing the block 1901 - 1950.

Figure 5.7. Relative proportion (%) of lexical items according to adjective class for Gor’kij, representing the block 1901 - 1950.

Figure 5.8. Relative proportions (%) of predicate adjectives according to class for Petruševskaja, representing the block 1951 - 2000.

Figure 5.9. Relative proportion (%) of lexical items according to adjective class for Petruševskaja, representing the block 1951 - 2000.

Figure 5.10. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Griboedov, representing the block 1801 - 1850.

Figure 5.11. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Turgenev, representing the block 1851 - 1900.

Figure 5.12. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Gor’kij, representing the block 1901 - 1950.

Figure 5.13. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Petruševskaja, representing the block 1951 - 2000.

Figure 5.14. Overall percentage changes in SF, LF-N and LF-I usage across the four time periods investigated.

Figure 5.15. Overall percentage changes in SF and combined LF (LF-N + LF-I) usage across the four time periods investigated.

Figure 5.16. Changes in SF, LF-N and LF-I usage for verbs where the whole corpus was searched (a total of ten different verbs).

Figure 5.17. Changes in SF, LF-N and LF-I usage for verbs where a subset of the corpus was searched.

Figure 5.18. Percentage change in occurrence with бывать (509 total examples; 49, 234, 67 and 139 in each time period).

Figure 5.19. Percentage change in occurrence with делаться (124 total examples; 15, 36, 16 and 87 in each time period).

Figure 5.20. Percentage change in occurrence with прибыть (21 total examples; 0, 4, 5 and 12 in each time period).

Figure 5.21. Percentage change in occurrence with представляться (14 total examples; 1, 7, 3 and 3 in each time period).

Figure 5.22. Percentage change in occurrence for казаться (426 total examples; 65, 19, 214 and 126 in each time period).
Figure 5.23. Percentage change in occurrence for оказаться (57 total examples; 13, 23, 7 and 14 in each time period).

Figure 5.24. Percentage change in occurrence for оказываться (77 total examples; 1, 45, 9 and 22 in each time period).

Figure 5.25. Percentage change in occurrence for оставаться (71 total examples; 26, 12, 24 and 9 in each time period).

Figure 5.26. Percentage change in occurrence for оставаться (98 total examples; 15, 42, 27 and 14 in each time period).

Figure 5.27. Percentage change in occurrence for стать (126 total examples; 36, 40, 26 and 24 in each time period).

Figure 5.28. Percentage change in occurrence for получиться (92 total examples; 0, 2, 15 and 75 in each time period).

Figure 5.29. Percentage change in occurrence for представляться (142 total examples; 3, 98, 16 and 25 in each time period).

Figure 5.30. Percentage change in occurrence for сделать (217 total examples; 51, 68, 71 and 27 in each time period).

Figure 5.31. Percentage change in occurrence for становиться (294 total examples; 26, 79, 34 and 155 in each time period).

Figure 5.32. Percentage change in occurrence for являться (294 total examples; 26, 79, 34 and 155 in each time period).

Figure 5.33. Percentage change in adjective occurrence for SF, LF-N and LF-I over time, based on the data presented in Table 5.1.

Figure 5.34. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective bolen (болен).

Figure 5.35. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective vinovat (виноват).

Figure 5.36. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective gotov (готов).

Figure 5.37. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective dalek (далек).

Figure 5.38. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective dovolen (дополнен).

Figure 5.39. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective dolzen (должен).
Figure 5.40. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective dostoin (достоин).

Figure 5.41. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective zdorov (здоров).

Figure 5.42. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective nameren (намерен).

Figure 5.43. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective nužen (нужен).

Figure 5.44. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective poxož (похож).

Figure 5.45. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective prav (прав).

Figure 5.46. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective svoboden (свободен).

Figure 5.47. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective sklonen (склонен).

Figure 5.48. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective soglasen (согласен).

Figure 5.49. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective sposoben (способен).

Figure 5.50. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective ščastliv (счастлив).
Acknowledgements

A thesis represents a piece of original research by an individual, in order to obtain a PhD. However hard a postgraduate student works, the input and support of a number of people is necessary in order for that student to complete successfully. To this end, there are several people who deserve my deepest gratitude.

Firstly, I would like to extend my thanks to "the three wise men" — my supervisors Prof Grev Corbett, Dr Dunstan Brown and Dr Matthew Baerman. Their patience, support and guidance have been invaluable in the completion of this thesis. Also, many thanks to the other members of the Surrey Morphology Group — Marina Chumakina, Andrew Hippisley, Anna Kibort, Bill Palmer, and Carole Tiberius, and my fellow PhD students within the Group — Claire Turner, Kasia Marchewska and Magda Flakowska. Angela Brooke and Catherine Slater, also PhD students in the Faculty of Arts and Human Sciences, have provided much needed tea (or beer), sympathy and understanding at various points throughout the process. Special thanks also to Dr Adrian Barentsen of the University of Amsterdam, both for provision of the Corpus used in this study, and for answering my questions on the decisions made during its production.

At the beginning of this research, although choosing to make a statistical analysis of the data, my knowledge was not what it was to become by the end. For this, I must thank Prof David Jones of The Queen’s University of Belfast and Dr Simon Wilkinson of Newcastle University.

As we began with three wise men, so we end with "the holy trinity" — my husband, Gary, for his unerring support and belief; my colleague on the short-term morphosyntactic change project, Alexander Krasovitsky, for being the perfect co-worker; and Dr John Pugh of Cardiff University, for much-needed help with the statistical analysis and occasional diversions into the Private Eye
crossword. All have shown limitless patience and have been supportive to an almost superhuman degree. To each of them, my thanks.

This PhD was funded by the Arts and Humanities Research Council (grant number RG/AN4375/APN18306), whose support is gratefully acknowledged.

Finally, I would like to dedicate this thesis to my husband and family, and to the memory of Thomas Asdell.
Chapter 1.

Introduction.
1.1 OVERVIEW AND CONTEXT

The PhD forms part of the AHRC-funded project “Short term morphosyntactic change in Russian” (grant number RG/AN4375/APN18306), which was conducted at the Surrey Morphology Group, University of Surrey, from September 2004 to May 2008. The stated aim of the project was to “provide a detailed picture of how a linguistic system can realign functions while leaving the forms intact” (where inflections remain constant, but their usage changes). All languages go through changes, and it was felt that not only would data from Russian help to understand the process, but also that Russian had undergone significant changes in a relatively short period of time in the sphere of morphosyntax, particularly in the following areas:

- The form of predicate adjectives
- Case government under negation
- Case government with numerals 2-4
- Agreement in conjoined noun phrases
- The case of predicate nominals
- Number in quantified expressions

The overall aim of the project was to “fill in the gaps” left by previous research into these areas, with the PhD thesis testing the final hypotheses on a “discrete system”\(^1\) in one particular area – the development of the predicate adjective in the period 1800–2000.

1.1.1 Overview of the thesis

The aim of his thesis is to quantify change in Russian predicate adjective usage over a two hundred year period, 1801 – 2000. This was done through examination of examples taken

\(^1\) Quotations in this paragraph are from the original grant proposal (grant number RG/AN4375/APN18306).
from the Barentsen Corpus, which were then divided according to pre-decided criteria and analysed statistically.

A predicate is a portion of the clause, excluding the subject, which makes a declaration or denial about the subject. It may occur with an adjective phrase (AP), noun phrase (NP) or prepositional phrase (PP). A copula verb — such as the verb *to be* — joins the subject to the predicate. In Russian, a predicate adjective can take one of three forms — the short form (SF), the long form nominative (LF-N) or the long form instrumental (LF-I). However, if the adjective occurs as part of a predicate noun phrase, the short form may not be used, as the adjective in this case will be standing in the attributive position. All three possibilities for the predicate adjective in Russian will agree in gender and number with their subject.

The changes examined in this study involve the choices made between the three possible adjectival forms. Further, the changes in the adjective form according to the verb type are examined by considering the differences which occur with subject noun phrases headed by different nouns — concrete, abstract, animate and inanimate. Concrete nouns are things which can be experienced in a sensorial, physical, way, while an abstract noun is something which cannot be experienced in such a manner. An animate noun is, for example, a person or an animal, specific or general, including proper nouns for names, and inanimate nouns relate to things which physically exist, such as a table or a chair, and may also incorporate abstract nouns, such as those given above. In addition, with regards to animacy and inanimacy, animate subjects will also automatically be concrete. With regard to the SF as a marker of a temporary state, it could reasonably be expected that this form would occur more frequently with animate subjects than inanimate, since animacy implies that an object is more easily changeable. However, Israeli (2007) suggests that the instrumental form is now employed in marking a temporary state, as the nominative is used in marking a quality which is inherent.

Finally, it has been suggested that case government properties of the verb will influence the choice of adjective form. Predicate adjectives occur with the copula and semi-copula verb, and semi-copulas typically govern the instrumental, so that semi-copula status would be expected to favour the LF-I. Verbs were further divided according to tense, aspect and
reflexivity. This was done in order to see if any of these favoured one form of the adjective over the others. If it was found to be the case that one form of the verb favoured either SF, LF-N or LF-I then the analysis conducted would enable the nature of this change to be characterised, particularly if this was spreading to other forms or if it was simply an anomalous occurrence.

This thesis focuses on developing and validating statistical methods that can be used to quantify language change, and to use these methods in conjunction with qualitative linguistic methods in order to better understand the process of language change in the context of corpus linguistics. Therefore, statistical analyses on how the frequency of adjectival forms change over the two hundred year time period which we are examining has been employed in investigating the Corpus.

The constructions under investigation are, and always have been, sensitive to both semantic and syntactic factors. However, as has been shown previously in a related study on case and predicate nouns (Krasovitsky et al., 2008), semantics played only a subsidiary role in the diachronic changes, supporting variation only at some stages, particularly in the first half of the 19th century. A semantically-based choice of nominative or instrumental was ultimately replaced by a single rule of instrumental case marking on predicate nouns. The shape of the shift from nominative to instrumental appeared to suggest that contemporary Russian is at an early stage in moving from a semantically-conditioned to a syntactically-determined model for predicate nouns. It is one aim of this study to determine whether similar trends in the adoption of the instrumental are observed using the Barentsen corpus.

Further, nouns were paired into concrete and abstract, and animate and inanimate, in order to examine what, if any, effect they might have on the choice of form. This aspect was deemed important since previous studies, as mentioned above, had shown that, in certain contexts, a SF was preferred for use with abstract objects. Finally, it has been suggested that the adjective form will be decided according to a particular component of the noun, such as concreteness or animacy, particularly as one of the main facets of the SF is that it is associated with temporality which, as pointed out above, may be associated with these semantic features.
In looking at individual adjectives, we have a number of issues to consider. First, we know from previous work and anecdotal evidence that there are a number of lexical items which now favour the SF, seemingly regardless of some of the factors deemed necessary for it. Previously, a group of adjectives has been identified as favouring the SF in the predicate position (i.e. Švedova, 1962). This group has been examined in the current study, and attention has been paid to those adjective classes that cannot, for example, take the SF. Also, searches were carried out regarding adjective class and for stress patterns, to see if either of these aspects could offer an explanation as to the perceived decline of the SF.

1.2 HISTORICAL LINGUISTICS

The understanding that linguistic change may be observed as it happens is only a recent phenomenon. For most of the twentieth century it was considered that language changed at such a slow rate that it could not be measured at the time; rather, it was thought that language changed could be measured only by comparing a point from the past with a point in the future, and mapping any changes. Indeed, Bloomfield (1933) stated that "the process of linguistic change has never been directly observed – we shall see that such observation, with our present facilities, is inconceivable." Such an opinion was held for the next thirty years with, for example, Hockett (1958) claiming that change could only be observed indirectly, via its consequences, and not directly.

However, Aitchison (1991: 18) commented that it was not so much that language change was unobservable, but knowing where to look, and what for, was of greater importance. In particular, researchers looked in the wrong direction as they uncritically employed methodological guidelines that were inappropriate for the purpose intended, providing oversimplified distinctions in the study of language. Such an example is the separation of synchronic (the study of the state of a language at a particular time) and diachronic (the study of language change) linguistics. A comparison of both methods was made by de Saussure, who commented that their separation was "absolute and allows no compromise". He employed the analogy of a tree trunk to compare the methods. A horizontal cut through the tree trunk allowed examination at a particular time (synchronic)
or a cut was made vertically, allowing the development over time to be addressed (diachronic). Usually, diachronic studies were composed of the results of synchronic studies, which were then compared to assess changes in the synchronic states over time.

While such an approach appears, superficially, to be logical, it does make the assumption that the data collected in synchronic studies is suitable for diachronic studies. It is entirely possible that they leave out the facts that are important for language change, thus inadvertently limiting their value to diachronic studies. Clearly, for example, language variation and language fuzziness are two such facts. Such issues impact further onto aspects of geographical, social and stylistic variations in a language. In addition, the nature of such changes is continuous in nature and therefore more amenable to analysis by suitable methods, such as the probabilistic approach to analysis of language change. The methods employed in this study will be further discussed in Chapter 2.

1.3 PREDICATES IN RUSSIAN

The predicate position in Russian allows three different forms of the adjective to occur – the short form (SF), the long form (nominative) (LF-N), and the long form (instrumental) (LF-I). Table 1 gives the forms for the adjective gotov ‘ready’ in the predicate position.

<table>
<thead>
<tr>
<th></th>
<th>SF</th>
<th>LF-N</th>
<th>LF-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc.</td>
<td>gotov</td>
<td>gotovj</td>
<td>gotovym</td>
</tr>
<tr>
<td>Fem.</td>
<td>gotova</td>
<td>gotovaja</td>
<td>gotovoj/oiu</td>
</tr>
<tr>
<td>Neut.</td>
<td>gotovo</td>
<td>gotovoe</td>
<td>gotovym</td>
</tr>
<tr>
<td>Pl.</td>
<td>gotovy</td>
<td>gotovye</td>
<td>gotovymi</td>
</tr>
</tbody>
</table>

In addition, the example below relates to the SF with a present tense verb, which illustrates that the SF is not only found with a zero copula in the present tense:
Further, examples of each of the three forms in the predicate position are shown below, where: 1a and 1b relate to the short form, 2a and 2b to the long form nominal, and 3a and 3b to the long form instrumental.

1b.

Нет, Ната́лья Петровна; вы нездоровы,

Net, Natal'ja Petrovna; vy nezhorov-y,

No, Natal'ja Petrovna 2PL unwell- SF.PL

посмотри на себя.

posmotr-ite na sebja

look-PRES.2PL at yourself

'No, Natalia Petrovna, you are unwell, look at yourself.'  (Turgenev)

(Note – vy is the polite plural form of "you" in Russian)

1c.

Я думаю, что торг здесь неуместен.

Ja duma-ju, cto torg zdes' neumesten.

1SG think-PRES.1SG that haggling here inappropriate.SF.MASC

'I think that haggling here is inappropriate.'  (Veller)
2a. Izba by-la tesn-aja, malen'k-aja, no čist-aja
hut be-PST-FEM cramped-NOM.FEM small NOM.FEM but clean-NOM.FEM

'The hut was cramped, small, but clean.' (Gor'kij)

2b. Boj budet ser'ezn-yj
Battle be.3SG.FUT severe-NOM.MASC

'The battle will be severe.' (Švarc)

3a. ... žizn' ee ne by-la takoj
life her not be-PST-FEM so

... her life was not so lonely, miserable.' (Gor'kij)

3b. Ja togda by-l malen'k-im
1SG then be-PST.MASC small - INST.MASC

'I was small then.' (Čexov)
There is a large body of literature given to describing the short form in syntactic terms, to the elements governing the choice of adjective after a copula verb (particularly бывать), to lexical differences between the long and short form, and so on. Corpus studies have been carried out previously (Gustavsson (1976), Benson (1954), Baženova (1993) for Russian since 1800, Grannes (1984) and Larsen (2002) for earlier periods) in various aspects of the predicate adjective in Russian, including its use in highly specialised texts (Baženova) and a diachronic study from Puškin to the 1950s (Benson). These works will be discussed in greater detail in the literature review (Section 1.4.2).

The aim of this research is to examine the development of the predicate adjective in Russian over the period 1800 – 2000. Major changes in the predicate adjective had occurred in Russian before this, leading to a major change in use between the long and short forms (Table 1.2).

Table 1.2. Comparison of the use of long and short form in prehistoric Slavic and modern Russian (modified from Larsen, 2002).

<table>
<thead>
<tr>
<th></th>
<th>Prehistoric Slavic</th>
<th>Modern Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attributive adjectives</strong></td>
<td>Long and short form</td>
<td>Long form only</td>
</tr>
<tr>
<td><strong>Predicate adjectives</strong></td>
<td>Short form only</td>
<td>Long and short form</td>
</tr>
</tbody>
</table>

As Larsen points out in Table 1.2, the long form spread at the expense of the short form, restricting the latter to the predicate position, and even there not as the exclusive form. This dramatic change happened, as Larsen describes in some detail, over a period of approximately 1000 years. The developments examined in this research cover a later – and much shorter – period. Nevertheless, it is proposed that the changes in the Russian predicate adjective have been no less dramatic.
1.3.1 The influence of literary style

Stylistics is described by Simpson as “a method of textual interpretation... interested in language as a function of texts in context, and it acknowledges that utterances... are produced in a time, a place, and in a cultural and cognitive context.” (2004: 3) Therefore, it is the study of the linguistic forms and techniques used by an author in order to – in the broadest sense – “tell a story”.

<table>
<thead>
<tr>
<th>Abstract storyline</th>
<th>Represented storyline</th>
<th>Domain in stylistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLOT</td>
<td>DISCOURSE</td>
<td>Textual medium</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sociolinguistic mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Characterisation 1: actions &amp; events</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Characterisation 2: points of view</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Textual structure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intertextuality</td>
</tr>
</tbody>
</table>

Table 1.3. A model of narrative structure, adapted from Simpson (2004, p.20). My underlining.

Table 1.3, above, illustrates a model of narrative structure. The points underlined in the column “Domain in stylistics” are the ones which most concern us in the present study. The textual medium can also be described as the genre, which for the purposes of this study has been divided into Fiction, Non-Fiction and Drama. The underlined points are those considered to be of the most interest in our analysis. The sociolinguistic mode applies to the social and historical setting of the text, both internal and external. Characterisation 1 describes the mechanisms of the plot, actions of the characters and the events in which they find themselves. Characterisation 2 is of particular interest, in that it describes the
mode of narration, such as first person. As Simpson states "The perspective through which a story is told constitutes an important stylistic dimension not only in prose fiction but also in many types of narrative text..." Of particular importance is the character of the narrator, be it first (homodiegetic) or third person (heterodiegetic) (second person occurs less frequently), as this can have a great effect on the question of animacy in the text.

The Barentsen Corpus contains nearly 1000 texts, written by a collection of over 100 authors, covering a variety of genres such as memoirs, novels, plays or journalism. This allows those using the Corpus for research to take a broad selection of different styles into consideration during their searches. However, this can in itself present a problem, as differing styles mean differing uses of language. As we have seen, previous literature on the topic of the Russian predicate adjective has suggested that the short form is considered to be “bookish” (along with the instrumental), while the long form is considered more "colloquial". This being the case, it would be reasonable to assume that non-fiction texts would have a higher percentage of short form adjectives than fiction. Works of fiction are written in a number of styles and registers, and the proportion of direct speech within a text can also affect this style. For example, a memoir written in the first person may not use such formal language as a history of Russia written in the third person.

The question of the differences in style will be examined in this thesis through the partitioning of texts not only in terms of the dates at which they were written, but also through genre, in as far as they will be headed “Fiction”, “Non-Fiction” and “Drama”. These divisions should allow us to examine the different levels of usage of the three forms of predicate adjective in varying contexts. However, this also presents a problem. Table 1.4, below, lists the short form adjective counts for a selection of non-fiction texts from the Corpus.
Table 1.4. Table of usage of SF adjectives in non-fiction texts.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Author</th>
<th>Animate</th>
<th></th>
<th>Inanimate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Masculine</td>
<td>Feminine</td>
<td>Neuter</td>
<td>Plural</td>
</tr>
<tr>
<td>1801 - 1850</td>
<td>Lažečnikov</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1801 - 1850</td>
<td>Glinka</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>1901 - 1950</td>
<td>Averčenko</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1901 - 1950</td>
<td>Ill and Petrov</td>
<td>35</td>
<td>18</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>1951 - 2000</td>
<td>Mandel'štam</td>
<td>45</td>
<td>27</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>1951 - 2000</td>
<td>Paušovskij</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1951 - 2000</td>
<td>Terc</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1951 - 2000</td>
<td>Pelevin</td>
<td>22</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

Note: There is a strong bias toward fiction texts (approximately 75% of all examples) in the first half (19th century) of the Corpus, meaning that any non-fiction analysis results will be skewed by the available data. Measures will be taken in the statistical analyses to deal with this.

Particular attention is drawn to two of the authors, Glinka and Mandel'štam, particularly in the “Animate” section of the table. The figures for Glinka show a count of 34 examples of short form adjectives with masculine animate nouns, but none with feminine animate nouns. Mandel'štam shows 45 for masculine animate and 27 for feminine animate nouns. If these two texts were taken in isolation, one could be forgiven for thinking that short form usage with feminine animate nouns did not become a possibility until the second half of the 20th century, particularly if the gender and animacy of the noun were considered important factors. However, it is important in this case to consider the nature of the two texts – Glinka’s is about the Napoleonic Wars, and Mandel'štam’s her memoirs. When we look at the “Inanimate” part of the table, we can see that there are more examples of Mandel'štam using the short form with masculine inanimate nouns than there are in Glinka’s work, but the two are almost equal (17 and 15 examples respectively) with feminine inanimate nouns. Again, the nature of the texts must be considered: Glinka, in writing about war, uses many items of military vocabulary, including vojna (war), linija (line), pozicija (position) and so on.
For example:

1a.

Позиция Бородинская была длинна и шершава
Pozicija Borodinskaja by-la dlinn-a i šeršav-a
Line-FEM Borodin-NOM-FEM be-PST-FEM long -SF.FEM and rough -SF.FEM

'The Borodin Line was long and rough'

The lack of feminine animate examples can be explained by the subject and historical placement of his work. By comparison, Mandel'štam's memoirs contain many references to friends and family, and the short form is often used to describe them:

1b.

Я уверена, что Бабель ходил к нему не из трусости, а из любопытства...  
Ja uveren-a, čto Babel' xodi-l k nemu ne iz trusosti, a iz ljubopylstva...  
1SG convinced-SF.FEM that Babel' go-PST.MASC to him.DAT not from cowardice, but from curiosity...

'I am confident, that Babel went to him not from cowardice, but from curiosity...'
1c.
Larisa was capable of much more...

1d.
And Tixonov is always loyal to himself and his vocation, which he serves.

1e.
And Fadeev was perceptive.

It might be expected that drama, made primarily of characters talking to each other, might contain a lower number of “bookish” short forms. However, it appears that this may not to be the case, as shown in Table 1.5, below.
Despite the fact that we might expect a more colloquial form of writing, drama also shows a substantial number of short forms for animate nouns. In this instance, it is the temporary element of the short form which is being employed by the author, particularly with such adjectives as gotov ‘ready’, zdorov ‘healthy’, ščastliv ‘happy’, bolen ‘ill’ and so on. However, a division of the analysis, along the lines of that performed above, would exclude such forms from the second count and may show very few short form adjectives at all.

Initially, the analysis of the Corpus was divided into Fiction, Non-Fiction and Drama in order to look at the differences between colloquial and bookish language. However, it soon became obvious that, in the case of some playwrights, this was not the case. For example, Gor’kij’s characters often use a declamatory style of speaking which could not really be described as razgovornaja reč’. However, drama does provide us with plenty of examples of the use and distribution of the predicate adjective with animate nouns, so the distinction has been maintained rather than amalgamating Drama into the general category of Fiction.

In depth examinations of the use of the predicate in works of literary fiction have been produced - The Function of the Predicate in the Fables of Krylov by H Hamburger (1981), for example, relating specifically to its stylistic use. Nichols (1981) points to the differences...
in style between Rasputin, writing in "colloquial" style, and Bulgakov's deliberately bookish language, and the differences in predicate usage to be found in this. However, a broad stylistic analysis of the texts lies beyond the scope of the present study but will, as far as possible, be taken into consideration.

1.4 PREVIOUS WORK

The Russian predicate adjective presents a number of problems for linguistics, in terms of morphology, syntax and semantics. One of the most active debates surrounds the status of the short form adjective and the so-called "category of state" (the second of these two issues will be dealt with in greater detail in Chapter 5). It is therefore not surprising that a substantial literature exists on the topic, covering these issues from a number of perspectives, including corpus studies.

1.4.1 Usage

There are a number of sections in Russian grammars given over to the use of the SF (i.e. Offord 1996, Timberlake 2004, among many others), and these largely agree on the usage and semantics of the SF, as opposed to the LF-N or LF-I in predicate position.

Švedova (1952) examined the use of the long and short form adjective in modern literary Russian, particularly the opposition of LF-Ns and SFs regarding grammatical and lexical conditions, and which factors favoured the LF-N or the SF. She found that the main differences in usage were previously stylistic, in that the SF was favoured in the more bookish, formal styles of writing, and the LF-N was favoured in the more colloquial, less formal style. However, she concluded that these divisions are diminishing. Firstly, there was a dispersal in grammatical and lexical form, between the SF and the LF-N. Secondly, this was followed by a gradual encroachment of the LF-N into syntactic territory once occupied by the SF, therefore weakening the position of the SF, as it could occupy no position other than the predicate. Švedova states that this is a continuing process. The SF
Is no longer used in clauses containing verbs of physical motion in fiction texts, and is also rarely used in the temporal sense. Since the SF lost so much ground syntactically, it became less flexible and was reduced to a relatively minor role, becoming a "category of state". In contrast, the LF-N has more freedom in where and how it can function, and is therefore growing in usage. Finally, Švedová stated that the traditional stylistic distinction between the SF in non-fiction and the LF-N in fiction is also declining, as the LF-I is also taking over in non-fiction, "bookish" texts. The system is still functioning, but the status of the SF is being weakened.

Mjasnikov (1970) examined the differences between the long and short forms. He determined that the long form was the predicate form most used in spoken Russian and has been steadily replacing the short form in this function, although no timeframe was given for this change. This was accompanied by an expansion of lexico-semantic meaning and syntactic properties of long form adjectives. Mjasnikov also commented that the short form in spoken Russian is insignificant, as it is made up from forms which have lost their links with the long form, such as rad, dolžen, nužen, and so on. He concluded that the long and short forms can be freely interchanged without change in meaning, and that the short form was not developing in spoken Russian but that its use and development was mainly restricted to the literary form.

Borras and Christian (1971) gave an overview of the Russian adjective as a whole, including the predicate. They claimed that the instrumental was widely used in the predicate position with a semi-copula verb – particularly in the present tense - and that "In this respect there is a marked difference between nineteenth and twentieth century usage." (p.79). They also stated that the instrumental form in this position took the meaning of a quality at a particular time, making it the logical choice to replace the short form. In terms of the status of the predicate and its usage, Borras and Christian believed that in sentences of the type Na nebe bylo pasmurno – The sky was dull, "dull" is an adverb, not a predicate adjective. They acknowledged that the "state is a temporary one and not necessarily characteristic of a person or object." (p.94) and that this was defined by other researchers as a "category of state". Category of state was used as a means of describing those instances where there was no overt subject, for example in the English phrase "It is
raining”, the “it” does not stand in place of a noun – the phrase merely describes a general state. Therefore, the argument would be that as no noun is being described, the use of an adjective of any form is inappropriate. The status of the adjective/adverb has been – and continues to be – debated in sentences without an overt subject.

Groen (1998) presents an overview which takes into account the major factors which determine or influence the choice between the use of short and long form adjectives in particular contexts. Firstly, the lexical meaning of the adjective stays the same in either the long or short form. Usually, the general meaning of the short form is that, in any given example of its usage, there are specific restrictions – these are usually temporal, spatial, conditional or subjective measures of a particular situation. However, the presence of such a restriction does not ensure that the adjective is of the short form. Groen highlighted situations where the short form’s use was required, which included the matter of reciprocity of the subject, the idea of oversize, and the expression of the amount or intensity associated with the quality mentioned. In particular, the short form found wide usage when the degree of something, or its intensity, were to be emphasised.

There is a significant lexical difference between the use of the long and short form, and a blurred boundary between the lexical meanings and polysemy. Groen uses the example of “distinguished, prominent” (LF-N) and “visible, conspicuous” (SF) and relates their meanings (i.e. visibility) with their abstract meanings (i.e. prominency) as a means of exemplifying the non-arbitrary nature of classifications, particularly at the boundaries of different groupings. Groen, however, does not make the distinction between “meaning” and “abstract meaning” absolutely clear.

1.4.2 Historical and corpus studies

Corpus studies have previously been used to look at the usage – and changes of usage – in the short form adjective. Two of the studies discussed in the following paragraphs – Benson’s thesis on the use of the predicate adjective from Pushkin to the mid-twentieth century, and Baženova’s paper (1993) on the use of short form adjectives in Russian
scientific texts – cover all or a large part of the period to be analysed in this thesis, and are both diachronic. Gustavsson’s work on predicate adjectives with byt' covers a shorter period.

Larsen’s study of the evolution of the long and short form adjectives in Russian (2002) takes data from four Chronicle texts in order to trace the development of the SF/LF-N distinction. She finds that, initially, the SF held the stronger position of the two, in that it could be used both predicatively and attributively, where the LF-N could be used only in the attributive position. There was also a distinction between those adjectives used in definite and those used in indefinite constructions, although the usage is ambiguous in such constructions. Over a period of approximately 1000 years, the roles went through a major change, although the definite/indefinite dichotomy remains unclear. Larsen also raises the issue of the definite/indefinite opposition and development actually being one of predication/attribution.

Grannes (1984) examined the change in short form adjectives and participles as facultative objective predicatives (those which modify a direct object – “facultative” is non-permanent, and “obligatory” is permanent) in 18th century Russian. He used a corpus to show a decline in the use of short forms (i.e. SF/LF) from 40.49 to 32.68%. In general, his corpus shows that the ratio between short forms and long forms is approximately 1:3. He compares his work to Glinkina’s (1968), whose approach includes long forms which have no corresponding short forms. Grannes found that, in the 18th century texts from which his corpus was composed, the use of short form adjectives and participles as FOPs (Facultative Objective Predicative) had declined substantially by comparison with the previous century. However, they were more widely used in 18th century literary language than had been previously supposed, and cannot be wholly attributed to the “high style prose of Russian Classicism” which saw substantial use of the short form up until the 19th century. It is proposed that, in essence, the use of short or long form appears to be influenced by a stylistic criterion, or “style-marker”, a claim which the author substantiates due to the absence of the short form in Karamzin.
Grannes also indicates that these findings help to establish a continuing downward trend for the SF, which had been stronger in the 16th and 17th centuries, as the 18th century ratio is approximately 1 (LF):1.5 (SF) which, he states, "represents a radical regress of the SFs."

Hence, the nature of the Corpus and its sub-sections (i.e. literary sources) have to be carefully considered and contextualised when corpus studies are conducted, including the range of source material and the range of the test time period.

Benson (1954) investigates the development of predicate adjective usage in Russian literary prose from Puškin’s time to the 1950s. He finds that a significant change in usage occurred during this period. The research material upon which his study is based is subdivided into three parts; the classical (1828 – 1880), middle (1880 – 1917) and Soviet (1917 – 1952) periods. Short predicate adjective usage was widespread during the classical period (81% usage), which Benson attributes to the retention of Karamzin’s "traditional" style by the majority of major writers (i.e. Puškin, Gogol, Turgenev). The middle period saw this usage drop from 81% to 63%, reflecting a stylistic variance and the elimination of what Benson perceives as outmoded syntactic features. The gradual increase of long-form adjective usage (from 12% in the classical period to 25% in the middle period) he attributes to the decreasing influence of the "aristocratic" style of classical authors and the vernacularisation of Russian literature. Instrumental usage rose from 7% in the classical period to 12% in the middle period. Interestingly, Benson finds that these changes continued into his "Soviet" period. However, while the trends show directional changes from each of his periods, they are less pronounced in the Soviet period. For example, short form usage decreases from 81% in the classical period, to 63% in the middle period, to 59% in the Soviet period. Long form (LF-N) usage increases from 12% in the classical period, to 25% in the middle period and then to 28% in the Soviet period. This is mirrored by an increase in Instrumental usage, which increases from 7% in the classical period, to 12% in the middle period and then to 13% in the Soviet period. In all cases, there is a large jump in form usage between the classical and middle periods, followed by – again, in all cases – an incremental increase from the middle to the Soviet periods. Statistical tests were not performed on these data to determine significant differences between occurrences in each of the time periods.
The subjective nature of the author’s “period” selection may influence the trends observed, and a realignment of these three periods (possibly mapped with different descriptors) may yield results of differing significance. In his analysis, Benson arbitrarily selects three strata into which his dataset is split. These are based on historical and sociological parameters, the most obvious of which are the early traditional style (to 1880) and the Soviet style (1917 onwards). His three strata are unequal in their length and possibly in their total number of literature sources – the latter being difficult to define as this information is not presented in Benson’s study, as percentages are used, rather than the total number of examples. His “middle” stratum is essentially undefined and may be attributed to a change in literary style that is not quantitative but empirical. However, the historical periods defined for the strata did mark logical points in Russian literary history at which changes could be expected to emerge. In essence, given the proximity of two numbers in each analysis, it may be argued that his dataset could be split into pre-1880 and post-1880 strata. For example, Benson suggests that short-form usage changes from 81% in his classical period, to 63% in his middle period and then to 59% in the Soviet period. His analysis of long form and instrumental usage also shows a similar distribution, with usage in both cases rising from the classical period and reaching a plateau in his middle and late periods. It is clear that short form usage is in decline; however, Benson’s data, particularly his use of three statistical strata, would suggest that the decline is less after 1880, and the decline observed may not even be significantly different in statistical terms. Benson’s analysis of a range of literature from Puškin to the middle of the 20th century shows that a significant change had occurred in the predicate adjective in Russian, and that this change is not complete. Benson concludes that the understanding of these changes allows the complications that arise in descriptive studies to be addressed.

Gustavsson’s (1976) work is an attempt to establish rules for the choice of adjectival form with the copula verb byt’ ‘be’, and looks in particular at contextual factors: in certain contexts, all three forms are possible (oxota byla udučna-udačnaja-udačnoj). Gustavsson aimed to establish rules for the choice of form, paying attention to contextual factors. The clauses investigated all have various forms of the copula byt’.
The study was based on searches of a corpus containing both fiction and non-fiction texts. The Corpus spanned the period 1964 – 1971 and encompassed 67 texts. There was an approximate balance between fiction and non-fiction. This is not a diachronic study, in that Gustavsson was not so much examining change in the use of the short form, but looking at the distinction between which form of the adjective – instrumental, long or short form – was favoured in various constructions using the copula byt’.

His hypothesis – as with others, including Benson – is that the short form is “bookish” or “oratorical”, and the long form typically colloquial or vernacular. To this end, he claims a clear distinction between the two styles in the choice of adjectival form. Gustavsson also states that, while the choice of form can have stylistic value, this is not the case in all contexts. Searches throughout his corpus found that both the instrumental and nominative long form adjectives were favoured in most cases, but the short form was typically used in the following:

1. phrases with ‘in’ plus locative case
2. non-fiction texts rather than in fiction
3. when the substantive in the previous context is known or given (the long form and instrumental form are used when this substantive is new or unknown).

Further, the short form was also used with abstract subjects, which have a higher percentage of both short forms and the instrumental than concrete and animate subjects. This combination is not surprising, as abstract nouns are typically “bookish”. The short form was also affected by word-class membership (as ěto – it is; vse - all, to – that, this; čto – what; the relative pronoun kotoryj – which, what; and the infinitive have a high percentage of short forms). Also there are certain groups of subjects which tend to use one form more than the others (for example, words denoting times of the day use the long form) and used with expressions of place – ‘in’ + locative has a high percentage of short form. Gustavsson considered the point thus: “Style surely plays an important part here, for subjects of this type are characteristic of bookish language, not colloquial language”.

22
Clauses in which the subject has one or more adjuncts have a higher percentage of short and instrumental forms than those without. The author suggests that the percentage usage increases with the grammatical complexity of the adjunct. Clauses where the subject has an adjunct are typical of a “bookish” style. Clauses where the subject is a substantive used in a preceding context exhibit a high percentage of long forms, i.e. if the information is previously known to both speaker/writer and recipient. They are usually used where the preceding context is new or unknown, and the short form used when it is given or known. Gustavsson indicated that "...the category definiteness seems to be reflected here in choice of form.” Therefore, Gustavsson argues, definiteness has more impact on the adjective form than the clause type.

The case of animate nouns seems surprising as – in the semantic sense – one of the primary functions of the short form is – supposedly – to denote a temporary state. So, while the long form seems logical for concrete nouns, why is this also the case with animates? The findings listed above, and especially those relating to animate nouns, will be re-examined through searches in the Barentsen Corpus in Chapter 2 to 5.

Baženova (1993) examined the long and short form as used in scientific texts from 1800-2000. The usage of both forms in literary Russian was defined and then compared with scientific usage, to see if the latter compiled. It was found that although the use of the short form decreased during this period, its occurrence in relation to long forms was still higher in scientific texts than in general language usage. The function and frequency of various semantic groups of short adjectives were studied, as it was considered that they would perform significant functions in specialist texts. The most significant difference between general language and scientific writing was found in the functioning of the short form as a marker of a continuous quality, which in other forms of literary Russian was usually conveyed by the long form.

The choice of texts in this case reflects the particular nature of scientific writing, in that there is a very strict academic convention for the authors of these articles, and that the language used can definitely be described as "bookish". Baženova is particularly
concerned with the necessity to express both precision and abstractness, and how these terms can be used in conjunction with the need for evaluative language. The Corpus used contained 58 texts from different branches of science (geology, physics, chemistry and so on). The analysis showed that there was a greater frequency of short forms in scientific texts (13.1%) than in both business (2.5%) and arts (1.9%) writing (these figures are given as comparison with long forms – therefore 86.9% in business writing would have been long forms, and so on).

One of the most interesting findings of Baženova’s research is the use of the short form in the case of “evaluative semantics” (1993, 139), for example “время неизменно” Vremja neizmenno – ‘Time is invariable’. In this instance, the use of the short form is shown to increase from around 26% in the 19th century, to 53% in the twentieth, doubling its frequency. This is at odds with the received wisdom that the form is declining, but must – in this instance – be attributed to the peculiarities of scientific writing. The author equates this with the disappearance of the short form from texts including “significant emotional evaluation” (possibly referring to psychology texts). Unfortunately, Baženova gives only percentages for her findings (and there are very few percentages given), making it impossible to run a separate statistical analysis or comparison with our own data. However, these findings cannot be considered insignificant, as they show the definite influence of written style in short form usage. This work is based on a very formal, prescriptive and academic use of language, which is being used for a specific purpose. Therefore, it would be inappropriate to use this one quantitative example to say that the short form is not decreasing – indeed, it is increasing. Although this example accounts for only a small number of adjectives within a highly specialist area of writing, it would be wrong to discount it completely.

Ueda (1992) used a small corpus of eleven texts, all memoirs, in his study, which found 1457 examples of predicate adjectives either with an overt form of byt’; or as part of a clause. Predicate adjectives in the present tense were not considered, nor were those with ēto, to, etc., those with specific suffixes that disallowed a short form, and subordinate clauses of the type “Uderţat’ ego bylo nevozmoţno” (“Keeping him back was impossible.”). The relatively small sample of works, along with the strict adherence as to what should and
should not be included, allows Ueda to analyse in great detail the status of the predicate adjective in clause-level constructions. However, in using such a restricted selection of resources, in that all were not only non-fiction, but of the same type of non-fiction, Ueda is unable to compare his findings across genres, or even to estimate the effect that third-person narration might have, as opposed to the first-person used within memoirs.

1.4.3 Semantics

Timberlake (1986) – examining only the long forms - suggested that the predicate complement can appear either in the instrumental or nominative cases, and that the choice between these cases is correlated with a difference in the interpretation of the state described by the predicate complement. He divided the LF and instrumental predicate complements into three categories – descriptive nominative, temporal instrumental and modal instrumental. The descriptive nominative is explained in terms of the state described being “independent of the particular circumstances of the narrated world” (p.139), and which has existed previous to the beginning of the action being described.

\[
\begin{align*}
& V o d n-o & \quad \text{utro} & \quad E l j u a r & \quad p r i s - a l & \quad k o & \quad m n e \\
& \text{In one-NEUT} & \quad \text{morning} & \quad \text{Eluard} & \quad \text{come-PST.MASC} & \quad \text{to} & \quad \text{me.DAT} \\
& r a s s t r o e n - y i, & \quad s k a z - a l, & \quad \text{čto} & \quad s & \quad Z a n & \quad - o m \\
& \text{distraught- SG.MASC} & \quad \text{say - PST.MASC that} & \quad \text{with} & \quad \text{Jean-DAT} \\
& J u g o & \quad p r i l d j u č i - a s' & \quad n e p r i j a t n o s t' & \quad \text{unpleasant} \\
& \text{Hugo happen-PST} & \quad \text{unpleasant} \\
\end{align*}
\]

‘One morning Eluard came to me distraught, and said that something unpleasant had happened to Jean Hugo.’

(Example from Řenurg 9.716, in Timberlake 1986, p.140)

The action here takes place “one morning”, however we can assume that Eluard was distraught before the action of the sentence takes place, and that he may have continued to be for some time afterwards. In such examples, Timberlake states that “to the extent that
it is clear that a given sentence is the onset of an episode [...] the preference for the nominative is quite strong and the instrumental is virtually excluded." (p.140)

The temporal and modal instrumental categories relate to, in the first case (temporal) a state beginning near to the time of the narration, and the second (modal) associated with negative adjectives or those which could be understood to be negative within a certain context. For example, the temporal instrumental:

Stepan, mesjac provaljav-šis’ v bol’nic -e; vernu -lsja zdorovym.
Stepan month loll -PST in hospital-LOC return-PST healthy-INST

‘Stepan, having lolled around the hospital for a month, returned healthy.

(Example from Stadnjuk 119, in Timberlake 1986, p.142)

The original circumstances (being ill) have been changed by subsequent circumstances (being in hospital) to reach the state which the adjective now describes (being healthy). This change in situation necessitates, according to Timberlake, the use of the instrumental. Finally, the modal instrumental:

Čerez peredn-ij kraj protivnika oni proš-li
Through front -NOM line enemy-GEN they pass-PST.PL
nezamečennym-y
unnoticed -NOM.PL

‘They passed through the front lines of the enemy unnoticed.

(Example from Birjukov 25, Timberlake p.140)

In this example, the state does not change – they were unnoticed throughout the action of the sentence. Timberlake goes one step further in claiming that the use of the instrumental
takes into account the fact that there may have been another outcome, where another form (nominative) may have been purely descriptive of that particular moment.

The differences in the two cases are that the nominative suggests a state where there is no departure from the expected state of affairs, whereas the instrumental does. Timberlake suggested that, in temporal terms, the state has typically held for "some time" prior to the narrated time. The nominative implies that the state is independent of the particular circumstances in its world – the state may or may not be valid under different circumstances. By contrast, the instrumental begins near the narrated time, at which time the matrix predicate occurs. This use of the instrumental suggests the beginning of the state, as well as a departure from the circumstances of the previous state. The modal instrumental shows a strong affinity for lexical adjectives that are, or can be implied to be, negated, i.e.; nevredimyj – unharmed and nezamečennyj - unnoticed. Essentially, it is suggested that the difference in the nominative and instrumental in the predicate complement construction examined lies in a different relationship between the actual reported state and the possible expected states. Timberlake further builds on the temporal theme in Flier and Brecht (1985), in which he again turns attention to the varying factors involved in the choice of the predicate, and how temporal, semantic, morphological and aspectual influences combine. In doing this, Timberlake attempts to find a model which will allow us to look at all possible outcomes from a given state, according to the choice of predicate form.

Petterson (1972) proposes that the difference between verbs and adjectives in Russian is not formal but is based on a syntactic distinction, almost suggesting the passive or active possession of a particular property – for example, "whiteness" in the following example:
Bel -et parus odinok-ij
White-PRES sail lonely -NOM
'A lonely sail is shining white.' (Lermontov)

Parus bel -yj
Sail white-NOM
'The sail is white.'
(Petterson 1972: 43)

For certain adjectives which have no short form, it is the case that they are derived from underlying non-activity verbal predicates, hence the absence of a short form. He also comments that the short form was used as if "in time" whereas the long form was used as it was timeless, or "outside time". Another distinction proposed between the long and short forms is that the former is concerned with the quality of an object, whereas the latter is concerned with its state. Further, the predicative use of a long form results in a connection between an object (or phenomenon) and the abstract notion of the object's quality, similar to its use in an attributive sense. The "property" beholden of the object may be temporary or permanent, a general property that is both permanent of and independent from the object in question. In the case of a short form, the property of the object is permanent or temporary, but in a predicative sense the property always belongs to the object, and it is always a property sought in the particular object — it is the essence of the object, and is inseparable from it. This difference is defined as the fundamental difference between the long and short forms of an adjective in the predicate, and may be viewed as a "morphological device" for expressing the difference between the active and non-active in an object.

1.4.4 Syntax and Agreement

Most work in this area has concentrated on the distinction between the long and short form adjective, again using a number of methods but with many being based on — or inspired by
Corbett (1979) has examined the question of predicate agreement in some detail. This work is split into two parts – the first deals with predicate agreement with unspecified noun phrases, and the second with predicate agreement with incompletely specified noun phrases. He argues that subjects which have no specification for gender or number are neuter by default, and that this “neutral” form must be distinguished from the SF. In syntactic terms, Corbett argues, the SF is losing ground to the LF, but in constructions requiring the “neutral” form, the seemingly SF form remains rather than being replaced.

The second part deals with alternative agreement in the predicate, along with “almost specified” forms (such as otec s synom – father with son), which is done by setting up a “squish” (“a continuum along which lexical items can be placed” Crystal (2003)) for the predicate. It was shown that agreement with forms such as kto - who could be handled within syntax, and that predicate agreement could be best understood if all types, not only verbal, were analysed together.

Corbett also looked (in an earlier paper – 1975) at sčítat nužnym constructions and proposed a “more complex” structure to allow for an analysis of instrumental usage. This is an area which had received little previous attention, but which Corbett points out Galkina-Fedoruk (1957) and Chvany (1970) had discussed previously. This was achieved by syntactically deriving the NP, and resulted in the observation that one could avoid nužnym being derived directly from the short form.

In Hierarchies, Targets and Controllers, Corbett dealt with the predicate hierarchy and its implications for Russian predicate agreement.

finite verb - participle - adjective - noun

(The Predicate Hierarchy, taken from Corbett 1983)
Corbett points out that the further to the right we move, so the chance of semantic agreement increases. However, the honorific "vy" in the Slavonic languages provides a problem, as it is also used for "you" plural. This is examined in great detail with a number of examples from the Slavonic languages, and the statistical evidence provides a strong argument for the hierarchy, showing that semantic agreement does increase as we progress to the right of the hierarchy. In combining target hierarchies (p.173) Corbett illustrates the relationship between predicate and agreement hierarchies, showing the point of contact as being the verb predicate. He shows that they can be successfully combined once certain differences between the two have been isolated, "controller factors counteract the effect of the Predicate Hierarchy, so that the different factors must first be isolated to demonstrate their individual effect." (p.174)

Babby's "Transformational Grammar of Russian Adjectives" (1975: 32) was based on Chomsky's Aspects Model (1965), which further developed the ideas in his earlier Syntactic Structures theory (Chomsky, 1957). He suggests that both the short form and the long form are surface structure categories, both coming from the same deep structure category V (verbal):

"The long form appears to function as the active participle of the short form... the passive participle has both a long form and a short form, i.e. behaves syntactically just like an adjective, while the active participle has only a long form; when the subject vy" 'you (polite)' refers to one person, the predicate long form must be in the singular, and in the short form must be in the plural."

Babby suggests that long form nominative adjectives are, in fact, short form adjectives which have been marked for case. Therefore, if we accept Babby's argument, that the short form is not declining at all, but merely being more usually marked for case, the question is raised of those adjectives that do not have short forms. Babby's argument would appear to suggest that certain adjectives could exist only once they have been marked for case. However, this would constitute the vast majority of adjectives, meaning that the rule itself would be the exception to the rule. In a later paper (1998), Babby
proposes a syntactic explanation of "the complementarity of LFs and SFs" in that they do not occur in the same positions and therefore "have different syntactic structures." He argues that a point made previously — that "SFs are never NP constituents while LFs are always NP-internal" was incorrect in the second part, which could not account for sentences such as:

```
On vernu -isja  dom -oj  golodn-yj
He return-PST  home-INSTR hungry -NOM
'He returned home hungry.'
```

The adjective in this sentence can no longer occur in the SF, meaning that the LF must take the position outside the NP.

Siegel (1976) attempts to analyse the Russian long and short forms within the framework of Montague Grammar, presenting the two forms as "members of two distinct underlying categories". Montague Grammar links semantic and syntactic considerations — "The syntax and semantics work side by side in such a way that any two expressions that are the same syntactically will be the same semantically, though not necessarily vice versa" (p.299).

Siegel points out that "Scientific laws and similar statements invariably contain short-form adjectives or verbs, and not long-form adjectives, in the predicate position" (which may give further support to Baženova’s findings). The example she gives, "Prostrantsvo beskonečno" — Space is infinite — would appear to be at odds with the previously stated use of the short form as being for temporary states. While the temporary/permanent dichotomy between short and long forms would be extremely useful in scientific writing, it appears that — in this example at least — the opposite is occurring. The short form is used to describe a permanent state — the endlessness of space. Also, in this particular example, "space" cannot be compared to anything else; it cannot be endless in comparison to another "space" elsewhere (as opposed to the example Siegel gives on p.297 of "studentka umna" — the student is intelligent in an absolute sense, and "studentka umnaja" — the student is intelligent compared to the other students).
Siegel attempts to find an explanation for how the long form can appear in both predicate and pronominal position, when short forms occur only in the predicate. She differentiates the long and short forms, describing the long form as nouns in inflection and agreement and in not taking infinitive complements. Short forms are, by comparison, like verbs in inflection, agreement and their ability to take infinitive complements. Short forms are categorised in the same manner as verbs and are said to follow verb rules. Further, long forms cannot appear in most imperatives with *byт’*, whereas short forms can, as the kind of *byт’* that appears with short forms is a transformationally inserted “dummy support” used for tense and mood. Bailyn (1994) used X-Bar theory to build a syntactic argument around Siegel’s work, appearing to justify the distinctions between the long and short forms. He combined syntactic and semantic considerations (following Montague grammar) to explain the differences between long and short form and, according to Toman’s (1996) review, raises a perennial problem in the mapping of syntax and semantics, “Either we differentiate adjectives in their predicative and attributive use as members of two different categories, or we do not differentiate these two uses and fail to provide for a consistent mapping between syntactic categories and semantic types.” (p.595) This question is central to the definition of the status of the Russian predicate adjective, as both syntactic and semantic aspects have to be considered.

Franks (1995) work is developed within the framework of Chomsky's Government and Binding Theory (1981). Franks examines the conditions affecting case agreement between primary and secondary predicates and their antecedents and discusses an account of the case that arises when agreement is blocked, and proposes that the solution to the latter issue be formulated in terms of control theory. Short form adjectives are formally caseless. As secondary predicates they are usually marginal and possibly also stylistically marked. Their usage also depends on the particular adjective of interest. Franks states that “a secondary predicate takes an adjunct predication of some item in the sentence, auxiliary to the main subject-predicate relation” and continues to give the three possibilities:
• agree with this NP in all pronominal features, including case
• agree with this NP in all features except case, receiving instead a default case by some other mechanism, or
• agree with this NP in all features except case but not receive a default case, appearing instead as a caseless short form (p. 221)

It is the instrumental which is the predicate adjective’s most common – and stylistically preferred – form.

Richardson (2001) states that case agreement suggests that the predicate adjective is non-eventive. The instrumental case suggests that the secondary predicate is eventive and that a discrete change of state occurred at or around the event time of the primary predicate (agreeing with Timberlake, 1986). Richardson suggests that case features in Russian may be eliminated from syntax and that case is linked to interpretable tense and aspect features.

Pereltsvaig (2005) presents an overview of the distinction between short and long form adjectives in Russian, and attempts to propose a lexical alternative to Bailyn’s syntactic account (Bailyn, 1994). It is proposed that long form adjectives are specified for case and features pertaining to animacy, which are not present in the short form, and that short forms are simply “shortcuts for bundles of primitive features”. Pereltsvaig reviews the properties of short and long forms in Russian and summarises the tests used to determine in which category an adjective should sit; meaning, distribution and morphology. The categorical features of lexical items are deemed insufficient to determine by themselves morphological and syntactic issues, so other properties – such as specification for particular morphological features – should be considered when categorising adjectives. These issues are contextualised by Pereltsvaig in the wider sense of examining diachronic change across Slavonic languages, although this change is described as “gradual” and no time period for this change is specified.
1.4.5 Summary of previous research

Previous research into Russian predicate adjectives is broadly divided into four areas – the syntactic nature of the predicate adjective, the semantic nature of the predicate adjective, predicate adjective usage in Russian and corpus studies. This substantial amount of research has produced a varied set of ideas and results, the majority of which will inform the present study and, in the particular case of the historical and corpus studies, provide a dataset against which to test and compare my own.

Through the examination of predicate adjective usage, most researchers (Švedova, Mjasnikov and others) have agreed that the SF is the formal, or “bookish” form, and the LF (both forms) more colloquial or vernacular. However, Švedova argued that the SF was declining and that the differences between it and the LF were diminishing. Mjasnikov stated that the SFs most used in spoken Russian (rad – glad, etc) were insignificant as they had lost – or were losing – their connection to their LF. He also argued that SF and LF could be interchanged and used in place of the other without any loss of meaning. Groen found that the SF still held in particular constructions, and was used particularly in the emphasis of degree or intensity. He also stressed the lexical differences between the SF and LF, showing that the forms were not interchangeable (as Mjasnikov had stated).

Larsen’s study of early Russian showed that the SF had originally been the predominant form, standing in both predicate and attributive positions, but that the LF had become stronger over time. Grannes noted a decline in the SF in the 18th century, by the end of which the ratio of SF to LF was approximately 1:3, and argued that this was due to stylistics. Benson’s diachronic study over a 150-year period also showed a decline in the SF, and argued that this was due again to stylistics and the vernacularisation of Russian. Gustavsson’s synchronic study was an attempt to establish the rules regarding predicate adjective choice in Russian. Again, the SF was identified as the “bookish” form, but he identified a number of areas in which the SF was the favoured choice. Baženova’s diachronic study of scientific texts over a 200-year period (1800-2000) found something of an inversion of the previous use of the SF as a marker of temporality, in that it was being
used for continuous quality (space is infinite). She also found that there was a higher percentage of SF usage in such texts, as opposed to other non-fiction texts.

The place of the present research in this context is two-fold: firstly, it will involve examples derived from a larger and more diverse corpus than has been used previously and, secondly, will involve statistical analysis designed to track the various elements affecting the morphosyntactic changes in the predicate adjective over the 200-year time period, and assess their effects on each other. To these ends, the present study seems a logical progression in the study of the Russian predicate adjective. In addition, this research will show the value of statistical analysis in measuring the stages of language change and will offer a method by which linguists can tell the point of progression of a particular change.
Chapter 2

Methodology
2.1 INTRODUCTION

2.1.1 Issues in Corpus Design

Rather than being simply a random collection of texts – perhaps even words – a corpus seeks to be representative of a language. Of course, such design is subjective and perhaps, in the statistical sense of random sampling, it may even be inappropriate for it to be anything other than a simple and random collection of texts. However, the content and design of a corpus really depends on what it is intended to represent – specific types of texts, such as plays, novels, scientific publications, or a general collation of works of all kind in a language. The representative nature of the Corpus therefore determines not only the nature of the research questions which are used to interrogate the Corpus, but the range and applicability of the findings of any such investigation. Therefore, the issue of representative samples relates as much to the content of the Corpus as to any mathematical or statistical treatment of the results, including sampling of the Corpus. It also limits the contextualisation available by comparison with previous studies which, while they may use the same methodologies to examine the Corpus, may be working with different corpora. Hence, comparisons with previous studies, as described in Section 1.4, and their findings should be discussed and compared in the knowledge that constraints may be imposed upon both their differences and commonalities.

Appropriate sampling of the Corpus is central in valid study design. This is discussed in Section 2.4. Proportional sampling, such as that used in opinion polling for political purposes, may not be wholly appropriate to corpus studies. Rather than the results proportional sampling would produce, corpus studies are interested in the range of linguistic variants that occur in a language, or in describing one variety of a language that is relative to another variety of the same language. In this case, it is therefore important that the Corpus covers all the varieties of the language. Even those documents that are seldom read by the vast majority of a population, such as legal documents or scientific research papers, should be included in some way in order to ensure that the Corpus is representative of all aspects of a language. This may also be extended to written and spoken forms of the language. Therefore, a corpus should seek to include the range of
linguistic variation present in a language but not the proportions of variation. This stratified approach, which should be characterised, is representative of categories that in some cases have been subjectively decided and which may not match apparently similar categories in other languages.

Hence, diversity is an important aspect of corpus design. Different lexical, grammatical and discourse features exist across different varieties of a language, and each register may have its own pattern of use. Any corpus, therefore, that is interested in variation of language of any type should consider the diversity of texts that it includes. Such diversity should also include register variation and subject matter. The latter is particularly important for lexicographic studies as word frequency may vary with the subject matter.

In Table 2.1, the numbers of examples in the Corpus, which have been used in this study, are listed by time period only. While this stratification is essential in the analysis of this study, only focusing on the total number of words is not always appropriate. In essence, the number of texts from different categories may need to be considered. Inclusion of small numbers of particular types of texts may skew the Corpus and any analysis thereof. However, it may also be argued that, if such texts do exist in such small numbers, their inclusion at such levels is valid and representative. The methods employed in this study are outlined in Section 2.3, and discussed thereafter. It should also be noted, as discussed in Chapters 3, 4 and 5, that a lack of examples returned from a search of the Corpus can be considered to be either unrepresentative due to low numbers, introducing skew or undue overemphasis to the results, or that such a return more realistically reflects the occurrences of particular items more widely in a language, assuming that the Corpus is itself a representative sample of that language. This latter case is important for lexicographic studies, which require large corpora due to the infrequent occurrence of words and collocations often returned in such studies. While this may be the case for lexicographic studies it should be noted that diversity is as important as size in many corpus studies.

Finally, other considerations may also affect corpus quality and relevance. These include the use of materials for which permission has been obtained – more importantly, it should
consider those materials not included for this reason. Consistent methods for collection and inclusion of data—such as consistent methods of transcription or, if relevant, transliteration, must be considered. While many considerations such as those described above may improve corpus design, it is imperative to appreciate that a researcher can only really compare like with like; different corpora have different compositions for different reasons and extrapolations between studies should carefully and fully consider the nature of the Corpus design—and its original intended purpose—before drawing conclusion from different studies, even studies using the same statistical methodology.

The above discussion relates to synchronic studies. In the case of diachronic studies, the parameter of time has to be considered in addition to size and diversity of register. Further, a paucity of historical texts may create difficulties for equality of sampling, and representative sampling, across all time periods. This also depends on the nature of the historical enquiry, which may be as simple as examining the works of one author, or examining one form (i.e. the novel) across a defined time, or it may be extremely complicated, investigating a series of authors and different forms across a wide time period, as in this study. The Helsinki Corpus (http://www.ling.helsinki.fi/uhlcs/) and ARCHER Corpus (now based at the University of Manchester - http://www.llc.manchester.ac.uk/research/projects/archer/archer3_1/) were designed for this purpose, as was the Barentsen Corpus, employed in this study. Such corpora may have structural and inclusion criteria; for example, the ARCHER Corpus contains at least ten texts in each fifty year period, which are mostly chosen at random.

Inclusion of registers is an issue, particularly for diachronic studies. For example, they may represent a wide range of the types of writing found in each historical period; they may be speech-based registers (court documents or plays); or they may include registers that exhibit a continuous history across all time periods, if indeed it is possible to find registers that do sit across all time periods; in cases where such registers change, such as the presentation of medical research from case-based to experimental-based works, the analysis of the Corpus should highlight what, in effect, requires subjective interpretation. A truly random corpus may include some or all of the above, and certainly contain elements of personal styles of communication (diaries, journals and letters, for example), fiction
prose, popular exposition such as newspaper prose, specialist registers such as those associated with legal, medical and scientific prose, speech-based registers (such as face-to-face communications, and monologues such as sermons). Some of these examples, such as sermons and prose from plays are often idealised written representations of speech and should be used carefully. However, a corpus may be designed for a specific purpose and therefore its construction might be more specific. As language changes over time, issues such as tagging might be difficult, as such practices may change. Therefore, where possible, a like-for-like comparison might be the ideal solution, but it may not necessarily be possible to fully achieve.

The method of analysis may also be considered to be subjective and cannot be assumed to be standardised. For example, computational methods may differ in their codification and as such they may search a corpus differently. The use of concordancing packages may be useful in standardising methodologies for corpus searching, particularly if a particular programme is widely used and established in the field. While writing one's own programmes is often more specific and suited to a particular study, optimising the output from that study, it is limited by the lack of transparency associated with such approaches, as programmes of this type often remain unpublished, allowing no one else to use such software. This limits the validation for such programmes and for the findings of particular studies.

2.2 HISTORICAL AND STYLISTIC INVESTIGATIONS

2.2.1 Stylistic and historical studies

Computational techniques have been applied to the study of language databases since the 1960s, initially to investigate the style of authors, genres and historical periods and thence to examine historical language databases, particularly diachronic text corpora, which have allowed a wide range of investigations, including examining stylistic (by focusing on a particular author) lexicographic, grammatical and discourse matters, to be carried out on their contents. Such progress from the previous methods of textual analysis was inevitable.
given the technological advances. Indeed, they continue to progress and develop rapidly, and as such the techniques available today should not be considered as a panacea to all the requirements of the field. Technological and methodological improvements will continue to be made, and they will keep ensuring that more rapid and streamlined methods of corpus analysis are available.

The value of corpora, and the increasingly rapid methods that permit their interrogation, is shown in the development and use of the Helsinki Corpus. Although developed as recently as 1984, well in excess of four hundred academic works have been published exploring the Corpus. Indeed, in another more recent measure, an internet search for this corpus returns over twenty thousand hits. Such a diachronic corpus is a valuable resource for many researchers, providing a large collection of varied texts from across a wide time range, a number of texts from different registers and the ability to undertake rapid, quantifiable and repeatable searches for a number of lexical and grammatical constructions.

Biber et al. (1998) outlined an example of how the Helsinki Corpus has been used. It considers the parallel development of modal and semi-modal verbs marking obligation/necessity. In such a semantic domain, two modal verbs — *must* and *should* — are employed as they mark a strong sense of personal obligation or logical necessity in the case of the former, and which are more weakly represented in the latter. For example:

**Marking obligation:**

(a) Now we *must* look to the next stage of development (News)

(b) You *shouldn’t* go to bed too early (Conversation)

**Marking logical necessity:**

(a) As Sherlock Holmes said, “When you have eliminated the impossible, whatever remains, however improbable, *must* be the truth” (Academic prose)

(b) He is able to specify a rule for what that amplitude *should be*. (Academic prose)
Semi-modals such as *have to*, *got to* and *need to* may be used to mark similar distinctions in meaning; in this case, in the strength of the sense of personal obligation associated with the above phrases, relative to weaker expressions which include *ought to* and *supposed to*:

Marking obligation:
(a) You *have* to protect your eyesight. (Fiction)
(b) I've *got to* tell him. (Fiction)
(c) I *need* to stay here. (Fiction)
(d) Maybe I *ought to* call them now. (Fiction)
(e) You were *supposed to* leave. (Fiction)

Marking logical necessity:
(a) We know that the transmitted beam *has* to have its polarisation in this perpendicular direction. (Academic prose)
(b) There *ought* to be little difficulty in teaching these things. (Academic prose)
(c) This is *supposed* to be a respectable hotel. (Fiction)

Modal verbs such as *must* and *should* have a long history in English, from before the tenth century. Semi-modal verbs marking obligation/necessity occur more recently, compared to modal verbs (ca. 1400). Historical duration of a verb does not indicate productivity, but extent of use can be readily determined by employing a corpus-based approach. In this case, Biber et al. examined frequency of modals and semi-modals in texts across four centuries. They used a range of sources, including the ARCHER Corpus, the Longmann-Lancaster Corpus and the British National Corpus. They plotted the relative frequency of modal and semi-modal occurrence over their chosen time periods for all three registers used. They found that, overall, modals were more common than semi-modals over the found hundred years of their study, but that the frequency was changing over time as semi-modals increased in popularity, being concomitant with a slight decrease in the occurrence of modals. Semi-modals had, by 1990, become as frequent as modals in news, slightly less in fiction texts but of greater frequency in conversation.
Adopting a lexicographical reading of these findings indicates that they do not follow the same patterns of use. The most common forms of modals have changed across the time periods of the study, for example, whereas the forms of semi-modals have remained reasonably constant over the same period. However, this must be taken in the context of the relative starting positions of modals and semi-modals, suggesting that such a change in semi-modal use might follow that observed for modal verb use. They were able to highlight specific usage, and suggested that this methodology could be developed to analyse the corpora in more detail by, for example, examining the specific contexts in which certain forms are used and mapping this context to frequency or by studying grammaticalisation.

2.2.2 The historical evolution of written and speech-based registers

Corpora can be used to track changes in various registers over time periods, allowing the analysis of a written language over time to assess, for example, literacy, or to quantify improved academic achievement. Such analyses provide an understanding of the changes to written language, and provide a good counterpoint for analysis of changes in speech. One issue in comparing the spoken and written forms is that the former is usually transcribed and is, in effect, not speech but speech-based and therefore an example of indirect evidence. This is expounded upon by Biber et al., (1998) who used the ARCHER Corpus to compare the changes in medical research articles and drama texts using four sampling intervals over a period of 350 years. These registers were chosen as they were generally regarded as opposites in style, the former being a formal written record, and the latter a fictional representation of various conversations (usually face-to-face, in the majority of examples) which provides an approximate representation of the spoken form. Comparison of two such registers therefore allowed the authors to compare, or infer, changes in the written and spoken registers across their chosen time period.

They compared two major dimensions of register variation, “Involved versus informational production” (including private verbs, present-tense verbs and second-person pronouns, for example) and “Non-impersonal versus impersonal style” (defined by passive constructions in both main and dependant clauses).
Their results showed that both registers were very similar in the seventeenth century, with literature in particular being characterised as impersonal compared to the latter parts of the study, but digressed thereafter with the dramatic prose becoming less impersonal and the medical text becoming more impersonal. No comparative statistical analysis was conducted across the time periods, so it is not possible to state, in a statistical sense, whether the changes observed are significant or not. In the case of drama texts, the twentieth-century texts have developed more of the involved characteristics and progressed to a near-absence of passive constructions. Medical prose progressed in the opposite direction, where information became increasingly dense and formal and increasingly passive in its constructions. This is in contrast to the colloquial style of medical texts from the seventeenth century, which were often letters written to editors for publication with little modification. A general contrast between medical writing of these eras, highlighted below with suitable examples, is the change from a style which could be understood by a non-medical reader, to a style which is explicitly technical and which requires substantial subject-specific knowledge to understand.

The first example is the title of a medical letter, written by Brown in 1685, entitled “A Remarkable Account of a Liver, appearing Glandulous to the Eye”, and appearing in the Philosophical Transactions of the Royal Society of London. Such a text is characterised by a combination of passive and involved examples, combining the use of very involved features with a vocabulary comprised of relatively dense, or verbose, nouns and other features. The use of the words “appearing” might suggest a degree of passive or subjective observation in the construction. Such historical writings were compared by Biber et al. to more recent works, such as that taken from the New England Journal of Medicine in 1985:

During the sequential activation of the complement components of the alternative pathway by a variety of surfaces, such as new cuprophane membranes or yeast-cell walls, peptides are split from the third (C3) and fifth (C5) components of compliment. These peptides are designated C3a and C5a, respectively, and the latter has the capacity to induce leukopenia and pulmonary leukostasis in vivo and to aggregate
granulocytes \textit{in vitro}. The carboxyl terminal amino acid arginine, of both peptides C3a and C5a, is rapidly cleaved by carboxypeptidase N of the blood, and the resultant stable peptides are designated C3a desArg and C5a desArg, respectively.

Comparisons of both the texts discussed, among many others, would simply not be possible without a properly constructed and representative diachronic corpus. Such a corpus clearly demonstrates the enormous changes that can occur over time. In the case of the two examples used above, it should be noted that the latter example is extremely technical and reflects not only the change in language but in the understanding and complexity of biomedical processes. One example of this is the abbreviation of arginine, a peptide, to the suffix –Arg; this is a common shorthand description of proteins and peptides whose use is specific to particular biomedical fields. Hence, while the Corpus seeks to characterise evolution of language it may be able to do this in the context of the evolution of the aspects of a register which are essentially external to the Corpus, a phenomenon that may both contextualise the findings of the Corpus in a field outside linguistics but which should be taken into account when comparing registers from different fields where rates of progress and evolution may differ markedly.

\textbf{2.2.3 The Barentsen Corpus}

The short term morphosyntactic change project uses a corpus approach, in order to be able to use similar data from a 200 year time period. For obvious reasons, the Corpus has had to be compiled using written texts, as any recorded, spoken language would only allow us to study examples from approximately 1920 onwards. It also allowed those compiling the Corpus to include texts of similar genres for different time periods, allowing a more accurate comparison of data. This is essential in the compilation process, as one of the four main characteristics of a modern corpus (listed by McEnery and Wilson 2006) is "sampling and representativeness", meaning that a corpus should include as wide a range
of authors, styles and genres as possible, which would be “maximally representative of the variety under consideration.”

This research used a corpus, provided by Dr Adrian Barentsen of the University of Amsterdam, as the raw material for the investigation. It was already machine readable and therefore ready to be searched using the appropriate program. The Corpus contains approximately 30 million tokens, which we initially divided into 20-year time slots. The number of texts contained within each of these time slots is shown in Table 2.1.

Ease of availability is an obvious factor in the Corpus containing more texts for the twentieth century, particularly for the last forty years. There is also a greater variety of texts, in that there are more examples of both drama and non-fiction in the second half of the Corpus.

The Barentsen Corpus is a corpus of texts collected over approximately twenty years. It was initially intended to provide a basis for collecting examples of temporal conjunctions, with particular attention to tense and aspect, for Dr Barentsen’s research. The latter, “Basic”, corpus, was designed specifically to study changes in temporal conjunctions over a 200 year period, and particularly “adhortative” expressions (see Barentsen 2003a and 2003b for his findings and analysis). The concept was to build a corpus of four roughly equal parts, divided into 50-year periods, with each subset containing approximately 1.5 million tokens. Barentsen included as wide a range of authors and styles as possible, although his choices were often limited by their availability in electronic form. The lack of nineteenth century non-fiction texts can be attributed to their lack of availability. The collection of the texts was helped by developments in technology allowing texts to be taken from the Internet (lib.ru.net), where earlier they would have required an OCR (Optical

2 Also listed are “finite size, machine-readable form and a standard reference”.

3 “The main factor was the availability in electronic form. Because most of the texts were gathered from text-collections on the Internet, there has been a kind of natural selection. Most of the earliest texts clearly belong to the literary “canon”. But I have made some special effort to include some representative authors, that are less known today. For instance, I scanned and OCR-ed some texts of Bestuzhev-Milinskij. When I had the possibility to choose, I tried to include comparable amounts of writers with rather different styles. I also tried to include in any subset at least some material from plays”. A.A. Barentsen, personal communication, 27.11.06.
Character Recognition) program or, as with the first example in the Corpus, would have been typed in manually.

Table 2.1. A breakdown of the number of texts contained within particular time periods in the Bartensen Corpus (note: total word count = 31,585,550).

<table>
<thead>
<tr>
<th>Period</th>
<th>No. of texts</th>
<th>Total for century</th>
<th>No. of words</th>
<th>Total for century</th>
</tr>
</thead>
<tbody>
<tr>
<td>1801-1820</td>
<td>24</td>
<td>896,758</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1821-1840</td>
<td>137</td>
<td>3,367,414</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1841-1860</td>
<td>88</td>
<td>2,070,487</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1861-1880</td>
<td>107</td>
<td>5,405,059</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1881-1900</td>
<td>78</td>
<td>2,084,380</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>434</td>
<td></td>
<td>13,824,098</td>
</tr>
<tr>
<td>1901-1920</td>
<td>39</td>
<td>643,589</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1921-1940</td>
<td>114</td>
<td>4,545,601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1941-1960</td>
<td>47</td>
<td>1,376,313</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1961-1980</td>
<td>135</td>
<td>5,278,639</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1981-2000</td>
<td>225</td>
<td>5,917,310</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>560</td>
<td></td>
<td>17,761,452</td>
</tr>
</tbody>
</table>

2.3 DERIVATION OF DATA FROM THE CORPUS

Data were taken from the Corpus using the Dialing Concordance program. The program allows the researcher to search for various morphological forms of defined phenomena, such as short form adjectives. The program searches the text, and gives the results in the context of the sentence in which they appear, allowing for analysis of the sentence structure as a whole. Several examples are shown below, in Figures 2.1 and 2.2, together with information about their location on the researcher’s filespace. These examples also show that the searches did not always return the exact data requested.
These results were then manually disambiguated, as the program would often return some forms which were not short form adjectives (e.g. adverbs, forms of irregular verbs, particularly in the past, and plural forms belonging to other word classes). Figure 2.2 illustrates this problem. The program would also return adjectives which were part of a predicate noun phrase, where the adjective itself was an attributive modifying the noun heading the phrase. Once disambiguated, the results could then be analysed in terms of sentence structure, noun class, adjective type, and so on. This process provided whole-
number, quantitative data, and the numerical results were analysed using a number of appropriate statistical tests.

2.4 STATISTICAL ANALYSIS

2.4.1 Populations and Sampling

Statistical analysis of the Corpus allows a systematic differentiation of linguistic change, within the parameters and limitations of the chosen statistical tests. In consideration of an appropriate method of statistical analysis the nature of the variables within the population and selection of suitable statistical samples are paramount.

The Corpus is described in Sections 2.1.1 and 2.2.3. An understanding of its structure is essential in selecting the correct mode of analysis. Effectively, in using a corpus, the study involves a defined method of sampling from a population of fixed size in order to extrapolate those results to the language as a whole (the basic tenet of statistical analysis). This is analogous to an opinion poll of one thousand people whose opinions are intended to reflect those of the country as a whole.

In discussing statistical analyses, notation (or terminology) is important. At this point, we need to define, and differentiate between, the terms population and sample. The population is considered to be the entire number of observations that constitute a particular group. The sample is a relatively small number of observations from a larger, defined, population. Therefore, using the above example of an opinion poll, the statistical population is the entire population of the country, and the sample is the number of people polled as being representative of this population.

In the absence of defined sample data, the mode of sampling is one of the most important parts of the statistical test and depends on, in this case, the structure of the Corpus. Therefore, appropriate selection of the texts from the Corpus is essential in avoiding bias.
and the drawing of subsequently inappropriate conclusions. For example, random sampling ensures that each member of the population has an equal chance of being sampled, and removes the possibility of bias from the sample. On the other hand, stratified sampling involves subdivision of the population into groups, often called strata. In the case of the Barentsen Corpus, this means that it is partitioned at fifty year intervals. Stratified sampling reduces natural bias in selection of the sample and is essential in a diachronic study. Therefore, a sample in any study should always involve results from more than one member of a group, if appropriate subdivisions exist. In this study, arbitrary fifty-year stratum have been utilised. It should be noted that, within each defined strata, the sampling is wholly random in nature. This does not mean that the selection is lacking in all order, but that every example – in this case, within the defined stratum – has an equal probability of being selected. The nature of the random sample is, in some ways, analogous to the nature of the population, in that there will always exist a probability that a level of bias, however small, has been introduced into the sample. Selection of the sampling frames – in this study, the fifty-year strata – have been chosen for a range of reasons, including the volume of data returned from the Corpus. In some cases, apparent in this, and subsequent, chapters, there exist very few examples for analysis. Twenty-year timeframes were considered, and have indeed in some cases been analysed in initial exploratory studies. However, overall, there were insufficient examples to allow appropriate statistical testing and as such fifty year samples were convenient as they provided strata that could potentially return meaningful results from a statistical analysis.

In this study, the sampling has been random to try to remove any bias in the sample at all. In discussing this, the nature of probability must be considered. In the most common example, the tossing of a coin presents us with two options – heads and tails. The probability that either heads or tails will be selected is based on the ratio of each different option, divided by the total number of options available. In the case of a coin toss, the probability of either heads of tails is 0.5, or 50%. To extend this, one might consider ten balls, numbered one to ten. The probability of any one ball being selected is one in ten, or 0.1 (10%). Further, if the first three balls were coloured red, the next three green, and the remaining four blue, the probability of selecting a red ball or a green ball would be three out of ten (0.3, or 30%), and the probability of selecting a blue ball would be four out of ten
(0.4, or 40%). As discussed later in this chapter, the frequency of selection is also an issue, particularly when comparing small (i.e. finite) or large (i.e. infinite) frequencies. The proportion of colours actually drawn, or heads or tails observed, may vary depending on the frequency; if the frequency is small then the proportion drawn may not reflect the occurrence of a particular colour in the population.

Another important point relates to the proportion of coloured balls, as opposed to the total number. In some ways, the latter is not relevant whereas the former is. For example, if a population of ten balls contained three red balls, this would be the same as a population of 3,000 red balls in a total population of 10,000. Woods et al (1986) expand such an example to consider a larger frequency of sampling — in their example, it relates to a number of people (42) selecting coloured discs up to 100 times each. In doing this, they show that, while the proportion of red discs is fixed at 30%, they see that the proportion of red discs drawn varies from 20 to 39%. However, the average for the whole study is 29.95%, very close to the actual number of red discs in the population. This demonstrates clearly how the nature of the sample, population and sampling frequency can all impact on the results of the study.

It is also important to consider the statistical independence of the sample. Usually, this means that one variable does not give information about another. However, this is not always the case and, in a study of this nature where SF, LF-1 and LF-N are being compared, and where there exist conditions that mean some particular forms cannot be adopted, independence of the data is not possible. This may impose some form of conditional probability, where the probability that an event will occur (i.e. in this study, that an SF might be selected) may be reflected in the event that other events have occurred (i.e. such as the classification of LF-I and LF-N).

Random sampling is not haphazard or without method. Rather, selection of a truly random sample is only achieved by following a defined protocol. For example, if we were to select three members of a population whose total size is eight then we must ensure that each member of the population has an equal chance of being selected. If we label each member of the population A to H, then we can calculate the total number of possible options
available, i.e. ABC, ABD, ABE, ABF, ABG, and so on. The total number of options is 56. In this case, there is a one in fifty-six probability that a particular three-digit combination will be chosen. While this clearly produces a wholly random selection, its problem is that, very quickly, the number of different options available grows exponentially, so that a sample of four taken from a population of fifty would yield approximately 250,000 possible combinations. This process is simplified by separating each member of the population, as opposed to all the possible options. A more rapid method is to use random number tables or the generation of random numbers by a calculator or via a software package. The latter method was employed in this study, so that fully random sampling from the results of corpus searches was possible.

Within the samples selected, there are several important parameters that we want to analyse. Of particular interest are the different types of texts within the Corpus, namely; fiction, non-fiction and drama texts. Where this is not possible, due to the texts not being available within the Corpus itself, additional texts have been used from the Russian National Corpus. This allows for any stylistic “eccentricities” that a particular author might have, and helps to avoid any subsequent skewing of the data. Again, this method of source selection should help to ensure randomness in sample selection.

Therefore, in order to analyse the Corpus in a statistically viable manner, it was decided at this stage to choose three texts (one of each of the types listed above) from each of the time periods for disambiguation and analysis. Within these bounds of strata and type of text the sampling is entirely random.

2.4.2 Statistical methods of analysis

The nature of the statistical test used in this study is dependent upon the structure and content of the Corpus. For example, the $\chi^2$ test is used as it is a test that examines frequencies, and not scale data. It is the aim of this study to determine frequency data for two categorical variables – time period and form. In other words, if the data being examined are quantised measurements and not continuous, then
ANOVA would be inappropriate, and other tests, such as the Friedman or $\chi^2$-squared, would be more suitable. In this case the $\chi^2$ test is chosen as the method of analysis as it is suitable to analyse the Corpus data for changes in form over the time periods measured. The $\chi^2$, or chi-squared test, is a non-parametric statistical test used to determine if a distribution of observed frequencies (in this case, the occurrence of SF, LF-I and LF-N) differs from theoretical, or expected frequencies (in this case, the null hypothesis that occurrence of form does not alter significantly over time). The test therefore determines the discrepancies between the expected and actual outcomes, and whether or not any such differences are statistically significant.

The $\chi^2$ test may also be used to determine whether, for example, changes between time periods, or changes between two different grammatical forms (i.e. perfective and imperfective) differ significantly within the same time frames. In the former case, for example, the null hypothesis proposed would be that the frequency of grammatical form (i.e. SF, LF-N or LF-I) would be evenly distributed over time periods. However, this test may be limited – as would a number of alternative statistical tests – if the size of the dataset is insufficient for a reasonable analysis to be carried out. For example, a $\chi^2$-squared test cannot be used if the expected number of observations in a particular group is less than five. This, for example, would be the case for the long form nominative and instrumental forms (see Chapter 3). However, in such cases, aggregation of the data into two forms instead of three may be possible, that is, comparison of the short form versus the aggregated data for long form (LF-N + LF-I). This would result in a 2 x 2 matrix (two time periods, two forms), and as such would require the use of Fisher's exact probability because of the reduced degree of freedom. In SPSS, the result for Fisher's exact test will appear automatically if you enter a 2 x 2 data matrix. Therefore, the $\chi^2$ test will be used in order to assess the statistical nature of the dataset.

In conjunction with the $\chi^2$ test the result for Cramer's V is also presented where it is appropriate. This statistic measures the strength of association, or interdependency, between two categorical variables (for example, time period and form, in the current study). While the $\chi^2$ test tells us whether or not a statistical relationship is significant, it does not
inform us of the degree of significance – this information is provided, post-test, by the Cramer’s V statistic. The result of Cramer’s V test is a number between 0 and 1; a result close to zero suggests little association between the variables, while a value close to one indicates a strong association.

The texts were divided into three types: fiction, non-fiction and drama, due to possible stylistic differences. Each of these was processed individually, and texts for each were chosen from each of the time periods. Therefore, change was measured both within the types, in terms of diachronic results, and also in comparison with the others.

2.4.3 Examples of Statistical Analysis from Initial Searches

The following tables and graphs represent some initial statistical analysis of the first data searches carried out for this research project. These were used primarily to test the usefulness of the various forms of analysis for this particular aspect of linguistic research. A group of fifteen adjectives was used in this initial study:

<table>
<thead>
<tr>
<th>Slavic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>болен</td>
<td>bold</td>
</tr>
<tr>
<td>виноват</td>
<td>guilty</td>
</tr>
<tr>
<td>готов</td>
<td>ready</td>
</tr>
<tr>
<td>далек</td>
<td>far</td>
</tr>
<tr>
<td>доволен</td>
<td>satisfied</td>
</tr>
<tr>
<td>должен</td>
<td>necessary/must</td>
</tr>
<tr>
<td>достоин</td>
<td>worthy</td>
</tr>
<tr>
<td>намерен</td>
<td>intend</td>
</tr>
<tr>
<td>нужен</td>
<td>necessary</td>
</tr>
<tr>
<td>похож</td>
<td>similar</td>
</tr>
<tr>
<td>прав</td>
<td>correct</td>
</tr>
<tr>
<td>свободен</td>
<td>free</td>
</tr>
<tr>
<td>склонен</td>
<td>inclined</td>
</tr>
<tr>
<td>согласен</td>
<td>agreeable</td>
</tr>
<tr>
<td>способен</td>
<td>capable</td>
</tr>
</tbody>
</table>

Where the Corpus allowed. Although fiction texts were available across the two centuries, non-fiction and – especially – drama were not always covered in every fifty-year slot in the Corpus.
Figure 2.3, below, shows the difference in the results when the group of fifteen adjectives usually found in the SF in the predicate position are removed from the dataset. The red squares show the original points and the black dots the amended ones.

![Scatterplot of short forms before and after removal of 15 adjectives](image)

Figure 2.3. Scatterplot of short forms before and after removal of 15 adjectives

We can see that, after taking the "fossilized" forms out of the equation, the SF continues to fall to the end of the twentieth century. However, the most interesting point is that, with the twelve adjectives in question, there appears to be a rise in SF usage, compared with the previous time period, suggesting that it is these particular SFs which are making the difference in terms of the changing occurrence of form over time. Also, by maintaining their status as SFs in the predicate position, these particular adjectives appear to be "buckling the trend" of being replaced by the long nominative or instrumental forms in this position. Performing a statistical test (using SPSS® v.14) on this data (for example, a two-way ANOVA test) shows us that time period is highly significant ($p = 0.003$) with regard to the first time period when it is compared to the other three time periods. This is shown on the graph by the drop between the first and second halves of the 19th century. In the amended data the continued decline in SF is very significant. The correlation coefficient, $R^2$ (adj.)
would be 100% for a perfect fit, whereas in Figure 2.4 it is 96.8%, which is an extremely high value (and an accurate) fit for what is real data collated from the Corpus.

![Fitted Line Plot](image)

**Figure 2.4.** A fitted line plot showing the decline of the SF over the time periods examined in this study.

This suggests that time is the most important factor in this change, rather than it being a matter of chance. It should also be noted that this is a preliminary investigation of the data, and that, overall, given the nature of the data, a number of statistical tests will be assessed for suitability with this data.

### 2.5 DISCUSSION

In considering the overall corpus, or available dataset, it may not always be essential to analyse every single data point; rather, a representative sample, or target sample, can be used which should, within particular confidence levels, be representative of the overall dataset - in essence, a representative sample population taken from a larger body (i.e. sampling from the Corpus in this study) should, within the defined limits of the study, be statistically representative of all possible values within the larger population. The removal
of insignificant or significant data can be observed by sequentially removing defined subsets (i.e. individual verbs) to see if they impact upon the overall nature of the collated data. This is a wholly qualitative process and would require sound linguistic reasons to so proceed. It also relates to the size of the overall dataset, and can be validated by removing every subset of data (in a “leave-one-out” approach”) which may be present in such small quantities that it does not affect the overall nature of the data.

An example of how a representative sample size is used in normal life is an opinion poll. Pollsters will question a representative sample of the population and produce an estimate, valid within certain boundaries of confidence, which reflects the overall opinions of a larger population. In essence, statistically relevant sampling can mean that a small sample can be used to reflect the trends of a larger population. A more pertinent linguistic example might relate to the vocabulary of 6 – 7 year old children (Woods et al., 1986). The variable employed in determining the vocabulary of this group is the score of a comprehension test for vocabulary size, and the population is all the words that the 6 – 7 year olds know, which give all possible values of this variable.

Sample size relies on two key parameters, the confidence interval and the confidence level, which must be understood when interpreting any data resulting from such analysis. The use of a point estimator, where a single point in the data is used to represent the overall population, is clearly limited as the use of a different point will invariably produce a different value. The confidence interval describes the range within which the data is accurate. For example, a confidence interval of 3 would suggest that the value obtained from a sample of the population is accurate in the following manner: if the value obtained was 34%, and the confidence interval was 3, then the result would be between 31 and 37, i.e., 34 ± 3. The confidence level is a widely used statistical parameter that provides an indication of how often the population is within the bounds of the confidence interval. For example, the commonly used 95% confidence level means that you have 95% certainty that your answer lies within the confidence interval. Similarly, a confidence interval of 99% means that you have 99% certainty in the data lying within the confidence interval. In essence, reporting of data sampled from a larger population normally requires that both the confidence interval and confidence level are stated.
Three factors will determine the confidence interval at a defined confidence level: the sample size, percentage and the size of the overall population. Selecting an efficient sample size will save time and resources when analysing a large population. The minimum sample size required to yield a representative sample can be determined from the population mean, \( \mu \). In this study, the population mean would represent the mean number of, for example, short forms in the dataset. This will differ from the sample mean, \( x \), and the size of this difference is essentially seen as an error between the sample and overall population. In such a case the margin of error (\( E \)) represents the maximum difference between the observed sample mean and the population mean, i.e., the difference between \( \mu \) and \( x \). Rearranging this formula, we can solve for the sample size necessary to produce results accurate to a specified confidence and margin of error:

\[
n = \left( \frac{z_{0.12} \sigma}{E} \right)^2
\]

(Equation 1)

where:

- \( z_{0.12} \) is the critical value, being the positive z value that is at the vertical boundary for the area of \( \alpha/2 \) in the right tail of the standard normal distribution;
- \( \sigma \) is the standard deviation of the population; in the case of this study that represents the distribution of SF and LF occurrence for each verb;
- and\n
\( n \) is the sample size.

In general, the mean of the sample will be similar to, if not the same as, the mean of the entire population. The latter, for reasons explained above, will never be estimated directly. However, the use of a sample, within certain confidence limits, allows us to estimate the mean within a particular, defined range. Overall, "on average", the results will be the same; that is to say, the mean of an infinitely large number of sample means would represent the mean of the population. This suggests that the sample mean is a consistent estimator of the population mean. Given the criteria for accuracy described above, the sample mean is
likely to be closer to the true mean if the sample size is larger, reducing the confidence limits in the results.

If you use a larger sample then you can have greater confidence that this sample will be truly representative of your overall population. Equally, a smaller sample will ensure that you may have a sample that is less representative of the overall population. This relationship is not linear, as the distribution of data is seldom linear in most cases. This means that doubling the sample size will not result in a halving of the confidence interval. The accuracy of the sample will be affected by the percentage of data that falls into a particular category. For example, if — in a survey — 99% of respondents say “yes” and 1% say “no”, the chances of error are remote, irrespective of sample size. This impacts differently on different types of data and is not directly relevant to the current study. Of substantial importance is the overall population size. In terms of probability, the overall size of the population is in most cases irrelevant. For example, the opinions of 1,000 voters can often accurately reflect the opinions of 50 million — sample sizes of approximately 1,000 are commonly used by electoral opinion pollsters. Population is only relevant when the population size is small. In these cases, such as the number of verbs which appear infrequently in the current analysis, the data may be unrepresentative. For example, comparison of the graphs (for example, Figures 5.18 to 5.32, on pages 192 to 202) clearly shows the impact that population size can have on data analysis and reliability. This means that, if less than one hundred examples are found for an example then the above figures suggest that any analysis of this data may be unreliable, as it may not contain enough data points to allow a representative analysis to occur. The figure of one hundred as “boundary” is a wholly qualitative estimate, based subjectively on interpretation of the data in this study, which appear in the above figures to hold as a general rule. Such a boundary may be different for another study. Those figures which have more than one hundred examples show a clear distribution of form and reasonable trends, as opposed to occasionally random data and trends where less than one hundred examples are present.

This, therefore, provides a significant issue with regard to the statistical analysis of the current study, which is addressed in Section 3.5. As a general point, it may be considered that quantitative statistical analysis of the Corpus at this point is not fully possible due to
the occurrence of examples returned from searches. Low returns from some searches would affect others, in that they may not permit a fully quantitative comparison of the data to be made. This matter is referred to in Chapters 3, 4 and 5. In addition, such analysis may only be representative of the type and nature of texts which comprise the Corpus, and the relevance of this to work outside the Corpus would be limited. Use of statistical methods for sampling is problematic with the Corpus dataset. Can, for example, a representative sample be found without knowing the size of the whole sample (for example, all SF occurrences in the Corpus)? The infrequency of occurrence would also suggest that few, if any, specific adjectives could be used individually as a representative sample. This may also depend on the linguistic nature of the “data” being examined, which may assume certain behaviour. Section 3.5 discusses this matter further in the context of examples that cannot adopt particular forms.

The categorical nature of language is a fundamental concept in its linguistic interpretation. Words, not numbers or statistics, are important. This has led linguists to classically consider language in discrete, quantised categories and not as a continuum. This further suggests little or no opportunity for grammatical gradients across categories and suggests absolute points of difference – and/or change – associated with a categorical approach. This has been the case for the categorisation of linguistic phenomena, such as morphological or phonological items. Bod et al (2003) commented that, while language is largely defined in a categorical sense, it is also increasingly seen as displaying evidence of continua and gradient behaviour. They further commented that psycholinguistic experiments demonstrated that speakers’ choice of words and sentences were very well predicted by the combined probabilities of their subparts.

The analysis of the data in this study would suggest the occurrence of a continuum, despite the arbitrary selection of discrete time periods for analysis. It is therefore difficult to use a generative approach to define an end period to capture the end of one linguistic occurrence, and to then observe the beginning of the next. This is clearly observed throughout the literature, where time periods are often chosen for statistical purposes, sociological purposes or socio-historical reasons (i.e. Benson, 1953). In this study, for example, the use of stratified sampling has been discussed previously. This has resulted in
the consideration of fifty and twenty year strata, with the use ultimately of fifty year strata due to the low frequency of some examples through the analysis. Within such a statistically valid model it may be difficult to ascertain endpoints. However, that is not the aim of this study – as the results shown in this, and other, chapters clearly demonstrate, the change in occurrence of the short form is not specific to a particular time period, and that different lexical items show different changes throughout the 200 year sampling period of this study. It would therefore be appropriate to consider a wholly gradient, and therefore non-categorical, approach to assessing morphological change. Indeed, it may be argued that such an approach is more revealing of language change as it considers the whole time period and not simply the end of one discrete part of a dataset and the start of another.

Probabilistic, or stochastic, analysis may describe a wide range of probabilistic linguistic phenomena, including morphological alternations, syntactic well-formedness and sociolinguistic variation, among others. The aim of probabilistic language modelling is to characterise the sample being analysed in terms of a probability distribution over all possible options available. The probabilistic approach may be employed in circumstances where, for example, several options exist for the arrangement of words into a sentence, in which case the probability of the most likely combination can be modelled. Further, it may be extended to characterising how similar or different two different languages are or, in the case of corpus linguistics, how accurately the data in the Corpus might represent the language (Stolcke, 1994). It is a technique that has also found application in fields such as speech recognition, has formed the basis of machine-based translation and spelling correction, among other applications (Chen, 1996), and which has been employed to develop understandings of language acquisition (Seidenberg, 1997) as part of its use in the wider field of cognitive sciences.

Probabilistic processing has recently been shown to have significance in human cognition (Bod et al., 2003; Chater et al., 2006; Chater and Manning, 2006; Oaksford and Chater, 2001). In the context of linguistics, it is clear that language changes over time, and that this echoes synchronically over time across different age groups. Zuraw (2003) suggests that probabilistic reasoning could explain the maintenance of lexical regularities over time. Further, Bengio et al. (2003) discussed a probabilistic approach to statistical language
modelling by developing a model that can learn the joint probability function of sequences of words in a language. This goal was complicated by the dimensionality of language, where different word sequences may have the same meaning, something which is extremely difficult to model as the output of the model is usually a single option. However, Bengio et al developed a model that learns a 'distributed representation for word usage', which allows each training set of their model to output more than one option by training the model to understand similar options for a sentence. Their model was able to learn a distributed representation, as opposed to a single representation, of each word. It was also able to estimate the probability of each sequence of words in a sentence. It was defined as an artificial neural network. This means that a sequence of words that the model has not met previously is given a high probability of occurrence if it is constructed from words that are similar to those words already encountered by the model. While their model presented a reasonable degree of accuracy, they used two relatively small corpora as the time taken to train the model was substantial.

Inherently, knowledge of language variation is a key part of individualistic language acquisition and competence. Indeed, it is suggested that production patterns of particular languages can vary widely across individuals (Bates and MacWhinney, 1987; Mendoza-Denton et al., 2003). This may be seen in the use of specific variants by individuals, both in terms of frequency and context. They further suggest that the knowledge of variation must involve knowledge of frequency.

It has been suggested that frequency is important in the probabilistic understanding of language (Bod et al., 2003). Frequent words are often recognised more quickly than less frequent words, and ambiguous words are often recognised or interpreted in terms of their most frequent meanings (MacDonald, 1994; Townsend and Bever, 2001; Jurafsky, 2003). Frequent words lead leniting changes and are more prone to reduction in speech (Mendoza-Denton et al., 2003). Frequent combinations of phonemes and structures are seen as being more grammatically correct than infrequent combinations (Pierrehumbert, 2003). Further, Baayen (2003) commented that the relative frequency of derived words and their bases affected the morphological decomposability of complex words. While the above are only a few examples of how frequency is both a probabilistic phenomena, and how it
affects language processes, it is important to demonstrate how frequency – an implicitly probabilistic part of language and language variation – can impact on morphological change in language. This can clearly exert an influence on the results of any corpus study, based on the words chosen and their frequency not only in the Corpus but in normal use. For example, if we take four examples from the list in Section 5.1; болен bolen – 'sick'; доволен dovolen – 'satisfied'; должен dolzen – 'necessary'; склонен sklonen – 'inclined'. These words are not chosen randomly, but reflect the extremes of occurrence in the Corpus. If frequency indeed plays a role in language change then it is reasonable to postulate either that infrequent words (should they appear often enough to warrant a statistically relevant analysis) would change less than frequent words (again, assuming a similar statistical relevance), or that greater frequency results in an increased stability and less change (Pagel et al., 2007). With the four examples chosen above, it is clear that neither situation occurs, and no clear trend can be discerned. In particular, reference should be made to Figures 5.34, 5.38, 5.39 and 5.47. These figures each show the change in occurrence of SF, LF-N and LF-I over the four time periods examined in this study. They show various trends, described in more detail in Chapter 5. However, the point of abstracting these examples based only on their frequency of occurrence in the Corpus (indicated by the number of examples returned from searches, relative to other examples listed in Section 5.1) is important in terms of the randomness of sampling and the absence of bias from such analysis.
Chapter 3

The influence of the verb in the choice of predicate adjective form.
3.1 INTRODUCTION

In the previous chapter we discussed the methodological issues in both corpus design and statistical analysis relevant to our purpose of determining the degree and nature of change in Russian predicate adjective forms over time. We now proceed to apply the methods discussed to determine the extent to which the choice of verb has an influence on the form of the predicate adjective.

The verb *to be* has either a copula or existential/locational function. In the example below, the use of the copula with the noun phrase (NP) *brilliant footballer* means that Paul is treated as belonging to the set of brilliant footballers, while the verb *be* with the prepositional phrase (PP) has a different function, namely to make an assertion about existence or location in a place. As is well known, this distinction has a number of grammatical effects in Russian, in particular in relation to negation.

Paul is a brilliant footballer. (NP)
Paul is in the stadium. (PP)

The verb *to be* can also be used in a copula function with adjectival phrases (AP).

Paul is short. (AP)

Semi-copula verbs, such as *to become* or *to seem*, because they can stand in place of the verb *be* in its copula function, can therefore be used with NPs and APs but not with PPs:

The castle seemed an impenetrable fortress. (NP)
The castle seemed formidable. (AP)
*The castle seemed in the eastern part of the county. (PP)*

In Russian, a predicate adjective may take one of three forms – the short form (SF), the long form nominative (LF-N) and the long form instrumental (LF-I). This is the case only for predicate adjectives – if the adjective occurs as part of a predicate noun phrase, the short
form may not be used, as the adjective in this case will be standing in the attributive position. All three possibilities for the predicate adjective in Russian will agree in gender and number with their subject. For example, in the case of gender:

**Kino by-lo nov-o /nov-oe /nov-ym**
Cinema be-PST.NEUT new-SF/new-NOM/new-INS
'The cinema was new' (SF/LF-N/LF-I)

**Komjuter by-l nov /nov -yj /nov-ym**
Computer be-PST.MASC new-SF/new-NOM/new-INS
'The computer was new' (SF/LF-N/LF-I)

**Mašina by-la nov-a /nov-aja /nov-oj**
Car be-PST new-SF/new-NOM/new-INS
'The car was new' (SF/LF-N/LF-I)

**Sobak-l by-li šumn-ye**
Dog -PL be-PST.PL noisy-PL
'The dogs were noisy'

The following are examples of predicate adjectives with copula and semi-copula verbs, taken from the Corpus:

**Состояние Литвинова было порядочное...**
**Sostojanie Litvinov-a by-l-o porjadočn -oe...**
**Status Litvinov-GEN be-PST-NEUT considerable-NOM.NEUT**
'Litvinov's fortune was considerable...' (Turgenev)
Тогда жена его делалась мрачною,

Then wife his become-PST-REFL sullen-FEM.INS

'Then his wife became sullen...'

(Somov)

Он должен прибыть с часу на час.

Ona dolžen priby-t’s čas-u na čas.

3SG.NOM should.SF.MASC arrive-INF from hour to hour

'He should arrive any moment now.'

(A Tolstoj)

Оно никогда не бывает спокойным,

Ono nikogda ne byva-et spokojnym,

3SG.NOM never not be-3SG.PRES peaceful-INS

море в наших краях.

more v naš-ix kra-ix.

sea in our-PL.LOC parts-PL.LOC

'The sea in these parts is never inclined to be peaceful.'

(Aksenov)
3.2 AIMS

The aim of this chapter is to analyse, by an appropriate statistical test, or combination of tests, the change in the occurrence of SF, LF-I and LF-N over a two hundred year period which is divided into four parts. As the copula byt’ allows us to examine all three predicate adjective forms, this provides a suitable starting point for the current study. From here, we will continue through predicate adjective usage with semi-copula verbs, full verbs and the zero copula (present tense of byt’). Further, we consider the potential influence of the aspect of the verb (perfective or imperfective), reflexive versus non-reflexive verbs, and tense. Through this, we may explore the role played by the verb type in the choice of predicate adjective form.

3.3 METHODS

In this chapter we examine the development of the predicate adjective according to its verb. The analysis is divided into the following groups:

**Verb**
- Copula
- Semi-copula
- Full verb
  - Reflexive (reflexive here refers to those verbs with the -sja ending denoting state, appearance or manner, rather than the so-called true reflexives, denoting an action performed by oneself, on oneself, such as бриться (brit'sja) – to shave)
  - Non-reflexive

**Aspect**
- Perfective
- Imperfective
Tense
Past
Present
Future
Infinitive (although tense in this case is marked on another verb, or is taken as present, for example in phrases such as "You need to be strong")

Imperative

Conditional

We first analyse the forms – SF, LF-N and LF-I – within each of these sub-groups, to track any change over the period 1800 – 2000. We then compare across the verb types, in order to examine the differences in adjective-type choice made within each. In doing this, we can show what influence the verb-type has in the choice of adjective form in the predicate construction.

For the copula verb *byt'*, which provides far more examples than any other predicate construction in the Corpus, the sample was taken from the following texts:

1801 – 1850:

Bestužev – Аммалат-бек (Ammalat-bek), Изменник (Izmennik), Замок Эйзен (Zamok Ejzen), Латник (Latnik)

Glinka – Письма русского офицера (*Pis’ma russkogo oficera*)

Puškin – Повести покойного Ивана Петровича Белкина (*Povesti pokojnogo Ivana Petroviča Belkina*), Капитанская дочка (*Kapitanskaja dochka*)

Durova – Кавалерист-девица (*Kavalerist-devica*)

Gogol’ – Мертве́ дерев (Mertvyye duši)

Nareznyj – Российский Жилбаз (*Rossijskiy Žilblaz*)
1851 – 1900:

Тургенев — Дым (Dym), Отцы и дети (Otci i deti)

Толстой — Анна Каренина (Anna Karenina)

Чехов — Три года (Tri goda), Супруга (Supruga), Белолобый (Belolobyj) Ариадна (Ariadna), Убийство (Ubijstvo), Анна на шее (Anna na šee), Дом с мезонином (Dom s mezolinom), Моя жизнь (Moja žizn'), Мужики (Mužiki), В родном углу (V rodnom uglu), Печенье (Poženje), На подводе (Na podvode), Шульц (Šul'c)

Куприн — Чудесный доктор (Čudesnyj doctor), Дознание (Doznanie), Ночлег (Nočleg), Олеся (Olesja), Штабс-капитан Рыбников (Štabs-kapitan Rybnikov), Синяя звезда (Sinjaja zvezda), Колесо времени роман (Koleso vremeni roman)

1901 – 1950:

Горкий — Мать (Mat'), Детство (Detstvo), На дне (Na dne)

Каверин — Два капитана (Dva kapitana)

Булгаков — Мастер и Маргарита (Master i Margarita)

Бунин — Темные аллеи (Temnye allei), Танька (Tan'ka)

Гилляровский — Москва и москвичи (Moskva i moskvichi)

Сергеев-Сенсий — Валя (Valja)

1951 – 2000:

Мандельштам — Воспоминания (Vospominanija)

Шаламов — Колымские рассказы (Kolymskie rasskazy)

Симонов — Живые и мертвые (Žive i mertvie)

Стругацкие — Пикник на обочине (Piknik na obočine), Понедельник начинается в субботу (Ponedel'nik načinaetsja v subbotu), Трудно быть богом (Trudno byt' bogom), Волны гасят ветер (Volny gasjat veter), Христолюди (Kristoljudi)
In order to examine any changes occurring in predicate adjectives with semi-copula verbs, we took the following list of 20 from *Akademicheskaja Grammatika* 1970, and searched the Corpus for examples:

<table>
<thead>
<tr>
<th>Russian Verb</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>бывать</td>
<td>to be, be present, frequent</td>
</tr>
<tr>
<td>делаться</td>
<td>to become, get, grow</td>
</tr>
<tr>
<td>сделать</td>
<td>to become, get, grow</td>
</tr>
<tr>
<td>казаться</td>
<td>to seem, appear (to be)</td>
</tr>
<tr>
<td>показаться</td>
<td>to appear</td>
</tr>
<tr>
<td>показываться</td>
<td>to appear</td>
</tr>
<tr>
<td>оказываться</td>
<td>to be found, proved (to be)</td>
</tr>
<tr>
<td>оказаться</td>
<td>to be found, proved (to be)</td>
</tr>
<tr>
<td>оставаться</td>
<td>to remain, stay</td>
</tr>
<tr>
<td>остатся</td>
<td>to remain, stay</td>
</tr>
<tr>
<td>получаться</td>
<td>to prove, to turn out (to be)</td>
</tr>
<tr>
<td>получиться</td>
<td>to prove, to turn out (to be)</td>
</tr>
<tr>
<td>представляться</td>
<td>to occur, seem</td>
</tr>
<tr>
<td>представиться</td>
<td>to occur, seem</td>
</tr>
<tr>
<td>прибывать</td>
<td>to arrive</td>
</tr>
<tr>
<td>прибыть</td>
<td>to arrive</td>
</tr>
<tr>
<td>становиться</td>
<td>to become, grow</td>
</tr>
<tr>
<td>стать</td>
<td>to become</td>
</tr>
<tr>
<td>явиться</td>
<td>to appear</td>
</tr>
<tr>
<td>явиться</td>
<td>to appear</td>
</tr>
</tbody>
</table>

As there were likely to be far fewer examples of this particular construction than had been the case with the copula verb *byt'*; it was decided that a search of the whole of the Barentsen Corpus used in this study would be necessary in order to acquire enough examples for analysis.

The paucity of examples available for the majority of these verbs rendered the distinctions between fiction, non-fiction and drama usage both impractical and impossible. Therefore, the statistics presented here take into account all three groups, but maintain the time periods in order to represent any change in usage.
Using the data for the verb *делаться* (делаться) graphical representations have been shown in two different ways, in order to illustrate what was done with the whole group of data, and to highlight the implications of these manipulations. The first (Figure 3.1) shows a graph plotted according to whole numbers, i.e. the actual number of examples found per time period. It should be noted that the y-axis is therefore a number scale (a count of the examples returned from the search) and not a percentage scale. The second (Figure 3.2) shows the numbers as percentages of the total predicate adjective count for each time period. The comparative value of each method is discussed in Chapter 2. In all cases comparisons are made between percentage occurrences for outputs of all searches.

![Figure 3.1. Change in occurrence with *делаться* delat'sja (154 total examples; 15, 36, 16 and 87 in each time period).](image-url)
A further graphical representation (Figure 3.3) shows the result of a linear regression analysis applied to the data shown in Figures 3.1 and 3.2 (for SF only). It can be seen that the blue diamonds, representing the distribution of occurrences of examples returned from the search of the Corpus presents a very different distribution of data compared to the red squares, which show the change in percentage occurrence. Very different regression coefficients are obtained – 0.04 and 0.99, respectively, indicating a poor fit in the case of the numerical data, and an excellent fit in the case of the percentage data. A high regression coefficient (0.99, when a “perfect fit” of points plotted to a regression line would be 1.0) indicates a clear linear change in the occurrence of the SF with delat'sja. This is discussed further in Chapter 5. However, the aim of illustrating these graphs and highlighting differences at this point is to clarify how the data is manipulated and how it is represented in this, and subsequent, chapters.
The type of analysis discussed above expands on how results from previous studies of corpora have been interpreted statistically. As discussed in Chapter 1, there have been previous corpus studies in Russian predicate adjectives — for example, by Benson (1954) and Gustavsson (1976). Benson (1954) examined the predicate adjective from Puškin to the mid-twentieth century in a diachronic study encompassing some 50 texts. Gustavsson carried out a synchronic study of the predicate adjective with *byt’*in texts spanning 7 years. Later, Baženova (1992) considered the use of predicate adjectives in scientific papers over a 200 year period (1800 – 2000).

A full description and validation of the methodology can be found in Chapter 2. A range of texts were searched using Dialing Concordance. The results were manually disambiguated and analysed according to form and tense of the verb. The results were then collated and analysed using an appropriate statistical test, or combination of tests. Initially, results were analysed by a two-way ANOVA, and subsequently changes in frequency of occurrence were analysed by the Chi-squared test. Statistical analysis was carried out using either MiniTab® (v.12) or SPSS® (v.14). In some cases, particularly for the Chi-squared test, where insufficient examples were returned from a search of the Corpus, data could not in
these cases be statistically analysed. An alternative was to collate together the data (i.e. 
LF-N + LF-I) if it was appropriate to do so, in order to provide sufficient examples for 
analysis of the general change from SF to LF. In some cases SPSS® automatically 
performed such corrections, in carrying out, for example a Fisher’s Test.

In carrying out Chi-squared tests in SPSS®, three variables were defined: time period 
(1801 – 1850, 1851 – 1900, 1901 – 1950 and 1951 – 2000), form (SF, LF-N and LF-I) and 
frequency. Labels were added to the column headers in order to make the results more 
readily comprehensible. For example, forms were coded as follows: 0 = SF, 1 = LF-N and 2 
= LF-I. Data was entered in three columns: time, period and frequency. The frequency is 
the number of examples returned from the search of the Corpus, and which have been 
classified after the process of manual disambiguation. Data was examined by weighting it 
by frequency, by using the “Weight Cases” function in SPSS®. The Chi-squared test was 
carried out by using the “Crosstabs” function in the Descriptive Statistics section in the data 
analysis menu. The time period was moved to the “Row” box, and the form to the “Column” 
box. The Chi-squared test was selected and the analysis conducted. Example outputs are 
shown later in this chapter.

The texts used in this study represent fiction (e.g. novels, short stories), non-fiction (e.g. 
memoirs, journalism, history) and drama, from four time periods – 1800 – 1850, 1851 – 
1900, 1901 – 1950 and 1951 – 2000. This study does not consider “direct speech” acts in 
texts as being representative of the spoken language. The grouping for drama is justified 
as being a very particular, and specialised, form of literature likely to exhibit different 
characteristics to other works of fiction.

3.4 RESULTS

Initial analysis of the data was carried out by ANOVA. This showed that clear, and 
statistically significant, trends were apparent in the occurrence of form over time. However, 
as discussed in Chapter 2, it was decided that both parametric and non-parametric 
statistical testing would be necessary in order to properly analyse the results, both in terms
of measuring variables in deciding predicate adjective form, and in attempting to track the rate of change across the two centuries. Therefore, the focus of the statistical analysis of the data collected for this study was on the use of the Chi-squared (χ²) test, as it is a test that examines frequencies, and not scale data.

Essentially, the analysis carried out compares the frequency of form in a time period. This means that the distribution between SF, LF-I and LF-N is compared between two time periods. These results are summarised in Table 3.1, and are represented graphically in Figures 3.4 to 3.14 (as percentage occurrence of forms). Differences, significant or not, are based on frequency and distribution across the two time periods. Measurement of significant differences is between the same forms but in different time periods, and not between the occurrences of different forms in the same time period. Statistical analysis was conducted with the null hypothesis being that different forms would be evenly distributed across each different time period. This means that a p-value greater than 0.05 would confirm this hypothesis – such cases are highlighted in bold in the rightmost column in Table 1. If the p-value for a comparison is greater than 0.05 this means that that particular comparison suggests that there is no significant difference between the data being compared; that is, it is evenly distributed between the two time periods. Conversely, p-values less than 0.05 indicate that the null hypothesis is rejected, and that the data being compared are significantly different.

For example, the first row in Table 3.1 shows quite clearly that the null hypothesis cannot be rejected and that there is no significant difference between the first two time periods. The first row, where the distribution of forms from 1801 to 1850 is compared to the distribution of forms from 1851 – 1900, returns a value of 0.402. This is greater than 0.05, and thus agrees with the null hypothesis – distribution is the same across these two time periods. Row two, where the distribution of forms from 1801 to 1850 is compared to the distribution of forms from 1901 – 1950 returns a p-value of 0.000 (to three decimal places). In this case, the null hypothesis is rejected and the results are significantly different from each other. This suggests that there is a significant difference in distribution of forms between these two time periods. This is illustrated graphically for all the examples later in this chapter where sufficient examples exist to allow a statistical comparison to be made.
Copula verbs have been analysed by including present tense zero copula examples. All comparisons are significantly different, indicating significant differences between distributions in all time periods, except when the first two time periods (1801 – 1850 vs. 1851 – 1900) are compared. Analysis without the present tense zero copula examples returned the same results. In all cases, when semi-copula verbs are examined, the null hypothesis is rejected – with one exception (the comparison of time periods 1901 – 1950 and 1951 – 2000). This means that different forms are not evenly distributed except when this final comparison is carried out, so the null hypothesis is accepted, where the $\chi^2$-squared test suggests that there is an even distribution between semi-copula verbs in the periods 1901 – 1950 and 1951 – 2000. In all other cases, there is a significantly different distribution of forms across the time periods ($p = 0.384$).

Exactly the same trend is observed for non-reflexive verbs as for copula and semi-copula verbs. The distribution of forms is significantly different in all cases, except when the time periods 1901 – 1950 and 1951 – 2000 are compared. In this case, the p-value is 0.757, greater than 0.05 and the null hypothesis is accepted, i.e., there is no significant difference between the distribution of forms across these time periods. This trend is also observed for the perfective and imperfective verbs, where, again, distribution of the forms is significantly different ($p < 0.05$ in all cases) except when the 1901 – 1950 and 1951 – 2000 time periods are compared ($p = 0.188$ and 0.361, for, respectively, perfective and imperfective verbs). For reflexive verbs, the situation is slightly different in that there are significant differences between all but two of the sets compared. In this case, the null hypothesis is accepted for comparisons between 1801 – 1850 and 1851 – 1900 ($p = 0.250$) and 1901 – 1950 and 1951 – 2000 ($p = 0.110$).

This means that the distributions of forms in the first two time periods is the same, as is the distribution of forms across the last two time periods. Distribution of forms across all other time periods is significantly different (i.e. $p < 0.05$ in all cases, and the null hypothesis (that the different forms would be evenly distributed across all time periods) is rejected. Examples of the past tense copula show significant differences in all cases, except where the first two time periods are compared ($p = 0.103$). For the past tense semi-copula
examples, distribution is significantly different across all time periods. The same trend was also identified for the present tense semi-copula. In addition to the $\chi^2$ test, in one case a low occurrence of examples (LF-N, 1801 – 1850: only two examples were found in the Corpus) meant that a Fisher exact test was carried out on a 2x2 matrix of SF vs. combined LF (LF-N + LF-I). The result, discussed in Chapter 2 and generated automatically by SPSS®, was that there was still a significant difference in the occurrence of SF and, in this case, combined LF across the time periods.

Due to a small number of examples analyses of the present tense zero copula, and the infinitive copula, were carried out by comparing SF with the combined LF, as described above and in Chapter 2. Examples were all in agreement with the null hypothesis, and no significant differences were observed with the data in either case.
Table 3.1. Statistical comparison, by χ-squared test, for predicate adjectives with copula and semi-copula verbs: comparison of frequency of occurrence across each time period, and the statistical significance of trends.

<table>
<thead>
<tr>
<th>Type</th>
<th>Time Period</th>
<th>Time</th>
<th>Comparison</th>
<th>χ-squared</th>
<th>Cramer's V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SF</td>
<td>LF-N</td>
<td>LF-1</td>
<td>p-value</td>
</tr>
<tr>
<td>Copula verbs</td>
<td>1801 - 1850</td>
<td>413</td>
<td>39</td>
<td>13</td>
<td>0.402</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>435</td>
<td>52</td>
<td>19</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>418</td>
<td>89</td>
<td>12</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>376</td>
<td>67</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>Semi-copula verbs</td>
<td>1801 - 1850</td>
<td>222</td>
<td>19</td>
<td>187</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>576</td>
<td>119</td>
<td>516</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>244</td>
<td>97</td>
<td>554</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>210</td>
<td>89</td>
<td>588</td>
<td>0.000</td>
</tr>
<tr>
<td>Non-reflexive verbs</td>
<td>1801 - 1850</td>
<td>83</td>
<td>2</td>
<td>21</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>230</td>
<td>57</td>
<td>22</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>93</td>
<td>40</td>
<td>49</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>80</td>
<td>54</td>
<td>61</td>
<td>0.000</td>
</tr>
<tr>
<td>Reflexive verbs</td>
<td>1801 - 1850</td>
<td>139</td>
<td>17</td>
<td>168</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>346</td>
<td>62</td>
<td>494</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>151</td>
<td>57</td>
<td>549</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>142</td>
<td>35</td>
<td>537</td>
<td>0.000</td>
</tr>
<tr>
<td>Past tense copula</td>
<td>1801 - 1850</td>
<td>248</td>
<td>30</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>245</td>
<td>39</td>
<td>6</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>237</td>
<td>79</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>210</td>
<td>58</td>
<td>29</td>
<td>0.000</td>
</tr>
<tr>
<td>Past tense semi-copula</td>
<td>1801 - 1850</td>
<td>158</td>
<td>15</td>
<td>113</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>350</td>
<td>72</td>
<td>305</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>281</td>
<td>83</td>
<td>197</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>202</td>
<td>50</td>
<td>390</td>
<td>0.000</td>
</tr>
<tr>
<td>Present tense semi-copula</td>
<td>1801 - 1850</td>
<td>40</td>
<td>2</td>
<td>17</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>107</td>
<td>39</td>
<td>54</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>37</td>
<td>16</td>
<td>50</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>46</td>
<td>43</td>
<td>146</td>
<td>0.000</td>
</tr>
<tr>
<td>Present tense zero copula*</td>
<td>1801 - 1850</td>
<td>121</td>
<td>5</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>138</td>
<td>10</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>161</td>
<td>5</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>132</td>
<td>8</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>Perfective</td>
<td>1801 - 1850</td>
<td>146</td>
<td>14</td>
<td>91</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>209</td>
<td>49</td>
<td>231</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>90</td>
<td>44</td>
<td>196</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>86</td>
<td>40</td>
<td>245</td>
<td>0.000</td>
</tr>
<tr>
<td>Imperfective</td>
<td>1801 - 1850</td>
<td>76</td>
<td>5</td>
<td>96</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>367</td>
<td>70</td>
<td>285</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>154</td>
<td>53</td>
<td>358</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>124</td>
<td>49</td>
<td>353</td>
<td>0.000</td>
</tr>
<tr>
<td>Infinitive copula</td>
<td>1801 - 1850</td>
<td>13</td>
<td>1</td>
<td>11</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>17</td>
<td>1</td>
<td>12</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>0.000</td>
</tr>
</tbody>
</table>

*Fisher's Exact Test used (2 x 2 matrix)
In the case of future tense semi-copula, infinitive semi-copula, verbal adverb semi-copula, imperative and imperative semi-copula, insufficient examples were available after searching the Corpus to allow statistical tests to be carried out. These results are summarised in Table 3.2. They are not represented graphically due to the lack of examples returned by searches. Again, the reader is referred to the Methodology section of this thesis (Chapter 2) where this subject is discussed in more detail. It is also discussed qualitatively, where a minimum total count of 100 examples is referred to – again, in a wholly qualitative sense – as providing a reasonable limit to allowing statistical testing of the dataset. This, again, is highlighted by the restrictions of the $\chi^2$ test, where a minimum count of 5 in each cell is required to perform a statistically valid analysis. The examples discussed herein did not occur in sufficient quantities in the Corpus to allow a combination of LF to be compared against SF. In most cases, insufficient examples were present to allow any statistical analysis via SPSS®. Therefore, it is difficult to draw any tangible conclusions from trends (if any) observed with these verbs. Such an analysis might only be possible with a larger corpus and a larger dataset from which to carry out an analysis similar to that conducted here.

Table 3.2. Results from corpus searches for types that were returned in insufficient numbers for statistical analysis to be carried out.

<table>
<thead>
<tr>
<th>Type</th>
<th>Time Period</th>
<th>SF</th>
<th>LF-N</th>
<th>LF-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future tense semi-copula</td>
<td>1801–1850</td>
<td>3</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Future tense copula</td>
<td>1801–1850</td>
<td>24</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Infinitive semi-copula</td>
<td>1801–1850</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>5</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>3</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>3</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Infinitive copula</td>
<td>1801–1850</td>
<td>13</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>17</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>6</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Verbal adverbs semi-copula</td>
<td>1801–1850</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Verbal adverbs copula</td>
<td>1801–1850</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Imperative copula</td>
<td>1801–1850</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Imperative semi-copula</td>
<td>1801–1850</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1851–1900</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1901–1950</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1951–2000</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
These results are reasonably clear in their interpretation. When comparisons are made across time periods, most of these analyses result in significant differences between the forms in these different time periods (Table 3.1). There are, however, exceptions where there is an even distribution across two time periods. Comparison of copula verbs indicates that significant differences are not observed when the first two time periods (1801 - 1850 and 1851 - 1900) are compared (p = 0.402). For semi-copula verbs, there is no significant difference in distribution between the third and fourth time periods (1901 - 1950 and 1951 - 2000; p = 0.384). The former point is true of reflexive verbs (p = 0.250) whereas the latter point is also true of both reflexive (p = 0.110), non-reflexive (p = 0.757), perfective (p = 0.188) and imperfective verbs (p = 0.361). In general, comparisons are only compliant with the null hypothesis when the first and second time periods, or the third and fourth time periods, are analysed. This would suggest that there is a certain degree of consistency in the results, and would suggest a certain degree of consistency in the data.

Clear trends can be observed in Figures 3.4 to 3.14. On each of the subsequent pages, appropriate pairs of graphs have been coupled for easier comparison. For example, Figure 3.4 and 3.5 show the different frequencies of form for copula (Figure 3.4) and semi-copula (Figure 3.5) verbs. Clear differences are apparent; whereas copula verbs are dominated by the SF, a clear rise in LF (both LF-N and LF-I) is apparent. While there is, comparatively, a greater occurrence of LF-N than LF-I, the former increases from the first time period (1801 - 1850) to the third (1901 - 1950), decreasing thereafter whereas the latter shows little change over the first three time period, but increases significantly in the fourth time period (1951 - 2000). It is also clear that the distribution, or frequency, of forms is different between each time period, confirming qualitatively the statistical findings summarised in Table 3.1. In Figure 3.5, not only is the SF seen to decline across the four time periods, but the LF, particularly LF-I, has a greater occurrence and, while LF-N increases slightly after the first time period, it plateaus, suggesting that the comparative changes between SF and LF are really those where the LF-I is replacing the SF in usage.
A similar trend is observed for non-reflexive and reflexive verbs, shown in Figures 3.6 and 3.7, respectively. Again, for non-reflexive verbs the SF initially dominates and is then seen to decrease. In this case, the increase in LF is due to an increase in both LF-N and LF-I,
unlike the changes observed overall for the copula and semi-copula verbs (Figures 3.4 and 3.5). However, changes in reflexive verbs are again based on an increase in LF-I only, at the expense of a decreasing frequency of SF.

Figure 3.6. Frequency of distribution of SF, LF-N and LF-I across the four time periods for non-reflexive verbs.
Figures 3.8 and 3.9 show, respectively, past tense copula and semi-copula verbs. In the case of past tense copula verbs the SF decreases slightly, but in this case it is the LF-N that increases substantially across the time periods. While LF-I increases significantly in the final time period, it makes a small overall contribution to the occurrence of these verbs. The picture is significantly more complex in Figure 3.9, where no clear trends are observed for the change of semi-copula verbs. The significant differences for these occurrences (listed in Table 3.1) are most likely due to the different distributions that can be observed in Figure 3.9. Specifically, LF-N increases from the first to the second time period, remains constant in the third, and then decreases again in the final time period. LF-I increases slightly over all four time periods, and SF shows a decrease across the first two time periods, and then a slight increase. Clearly, there is no similar frequency of distribution in these particular results and the distribution of these forms within the Corpus is clearly inconsistent. One clear difference between past tense copula and semi-copula verbs is in the occurrence of LF-I, which is almost non-existent for past tense copula verbs, with 14 examples being found (Table 3.1) compared to past tense semi-copula verbs, where over one thousand examples were returned after a search of the Corpus.
Figure 3.8. Frequency of distribution of SF, LF-N and LF-I across the four time periods for past tense copula verbs.

Figure 3.9. Frequency of distribution of SF, LF-N and LF-I across the four time periods for past tense semi copula verbs.
For present tense zero semi-copula (Figure 3.10) and present tense copula (Figure 3.11) comparisons are difficult as the latter did not have any examples of LF-I after searching the database (as was to be expected). Further, the SF and LF-N remain reasonably constant and no trends are apparent. This is seen in the results of the $\chi^2$ test, where the results are in agreement with the null hypothesis in most cases. The present tense semi-copula (Figure 3.10), by comparison shows clear trends. The SF decreases significantly across the four time periods, LF-N increases to a plateau and then tails off, and LF-I again shows a large increase across the time periods. Similar trends are observed for the perfective (Figure 3.12) and imperfective verbs (Figure 3.13) although the trend for LF-I is not as clear, initially decreasing before increasing again – this is mirrored by an initial increase, between the first two time periods, in the SF, which decreases thereafter. Infinitive copula verbs (Figure 3.14) show an increase in LF-N up to the third time period, which is followed by a sharp decrease. While SF and LF-I decrease and increase, respectively, these trends are neither consistent nor clear. This may be due to the small number of examples returned from searches of the Corpus (Figure 3.4), such as the example below:

3e.  
Bое будеt готово к утру  
Vсе бe-3SG.FUT готов-o к утру-DAT  
All 'Everything will be ready by morning.'

(Dovlatov)
Figure 3.10. Frequency of distribution of SF, LF-N and LF-I across the four time periods for present tense semi copula verbs.

Figure 3.11. Frequency of distribution of SF, LF-N and LF-I across the four time periods for present tense zero copula verbs. Note: examples of LF-I cannot occur with present tense copula byf'.
Figure 3.12. Frequency of distribution of SF, LF-N and LF-I across the four time periods for perfective verbs.

Figure 3.13. Frequency of distribution of SF, LF-N and LF-I across the four time periods for imperfective verbs.
Figure 3.4 shows the distribution of the three adjectival forms with the copula byt' across the four time periods for 1801 – 2000. This figure indicates that there is hardly any decline at all in SF usage, with the trend in LF-N showing an increase up until the third time period before falling off slightly in the fourth. The LF-I barely registers until the final time period, at which point it begins to catch up with the LF-N. These trends suggest that predicate usage with the copula byt' favours the SF throughout the periods studied. In the case of these particular examples, this would seemingly contradict predictions of the demise of the SF; while there is a decline in SF usage it is from approximately 85% to 75%. Therefore, we need to examine other conditions under which the predicate adjective occurs in order to find those under which the SF decreases. We begin with semi-copula verbs, shown in Figure 3.5. Here, we see that the SF is not as dominant as under the previous condition, and shows a steady decline throughout the four time periods; in fact, its usage decreases by more than half. The LF-I is the form of choice, with the LF-N hovering at 10% after 1851. Many of the semi-copula verbs examined take an instrumental object, and while a predicate adjective could in no way be described in such terms, there could be a possibility
that the case normally used by the verbs in question has been extended to predicate adjective constructions.

None of these match the description of a true reflexive verb, i.e. one whose semantic subject and object are the same. Again, many reflexive verbs take the instrumental, therefore it is reasonable to assume that this is being extended to the SF, and that such an assumption would constitute logical usage. While this may be the case, it is clear from Figure 3.6, for example, that SFs are again declining, but no clear form emerges as the predominant choice—SF, LF-N and LF-1 are more or less equal (especially when one considers the boundaries of error, ± 5%). Thus, for the final time period we may assume that something other than the verb decides the adjectival form.

The lack of clarity presented in specific, rather than overall trends, is shown by the reflexives (Figure 3.7) where we find that they have a very clear trend towards the LF-1 (indeed, they have a clear and increasing tendency to adopt LF-1 throughout all the time periods examined), with SF declining until reaching a plateau in the 20th century, the LF-N showing very little (if any) change, and the LF-1 accounting for almost 80% of the examples analysed. Therefore, we can conclude that a reflexive verb is most likely to occur with an LF-1. We may also be able to infer that the reflexive verb preferentially, and increasingly, adopts LF-1. In Figure 3.8, above, we do not see much difference in past tense as opposed to all examples with copula verbs, suggesting that tense, in this instance, is not a deciding factor in adjective form.

Comparisons were also made between verb types in each of the four time periods. These results are summarised in Table 3.3, and shown graphically in Figures 3.15 to 3.34. This table, below, shows comparisons between, for example, copula and semi-copula verbs, and the occurrences of each form are shown next to each other in the “Form” columns. These comparisons indicate that there is, in almost all the cases analysed, a significant difference between these pairs across all time periods. The two cases where no significant difference is observed (i.e. the results are in agreement with the null hypothesis) are where infinitive copula is compared with infinitive semi copula for the time period 1951 – 2000 (p = 0.159), and where perfective is compared with imperfective for the time period 1901 – 1950.
(p = 0.171). Clearly, such cases are exceptions to the trend observed in the remainder of the data – that is, significant differences are apparent in virtually all the comparisons made. Inspection of the data would appear to suggest that significant differences are predominately the results of the differences observed for the LF-I form and, in the comparison of the infinitive copula with the infinitive semi copula (1951 – 2000) the differences are most likely due to the small sample size found in the Corpus.

Table 3.3. Statistical comparison, by χ-squared test, of pairs of verbs, in each of the time periods, and the statistical significance of trends.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Form</th>
<th>χ-squared</th>
<th>Cramer's V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time Period</td>
<td>SF</td>
<td>LF-N</td>
</tr>
<tr>
<td>Copula (left) vs. semi-copula verbs (right)</td>
<td>1801 - 1850</td>
<td>413</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>435</td>
<td>576</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>418</td>
<td>244</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>376</td>
<td>210</td>
</tr>
<tr>
<td>Non-reflexive verbs (left) vs. Reflexive verbs (right)</td>
<td>1801 - 1850</td>
<td>83</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>230</td>
<td>346</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>93</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>80</td>
<td>142</td>
</tr>
<tr>
<td>Past tense copula (left) vs. past tense semi copula (right)</td>
<td>1801 - 1850</td>
<td>248</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>245</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>237</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>217</td>
<td>302</td>
</tr>
<tr>
<td>Present tense zero copula (left) vs. present tense semi copula (right)</td>
<td>1801 - 1850</td>
<td>121</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>138</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>161</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>132</td>
<td>46</td>
</tr>
<tr>
<td>Infinitive copula (left) vs. infinitive semi copula (right)</td>
<td>1801 - 1850</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Perfective (left) vs. imperfective (right)</td>
<td>1801 - 1850</td>
<td>146</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>209</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>90</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>86</td>
<td>124</td>
</tr>
</tbody>
</table>

*indicates analysis of SF vs. (LF-N + LF-I), Fisher's Exact Test, due to small sample size.

Examination of the graphs indicates several interesting changes to the occurrence of forms across the time periods. For example, in the comparison of copula and semi copula verbs, the semi copula SF can be seen to decrease significantly, whereas its copula counterpart decreases only slightly, remaining reasonably static across the sampling intervals. Copula LF-N increases over time, whereas semi copula LF-N appears to only increase from the first time period, whereafter it reaches a plateau. By contrast, semi copula LF-I appears to increase in the second half of the sampling period, whereas copula LF-I – whose occurrence is limited across the dataset – demonstrates a rise in the last time period only.
Comparisons between reflexive and non-reflexive verbs show a similar distribution, when both form and type are qualitatively compared (Figure 3.19 to 3.22). For both types, SF decreases with time, LF-N (non-reflexive) increases significantly in its occurrence over time, whereas LF-N (reflexive) does not appear to show any significant trends. LF-I shows a substantial increase in occurrence of the reflexive, and a small increase in occurrence of the non-reflexive, over time.

Figure 3.15. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1801 - 1850.

Figure 3.16. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1851 - 1900.
Figure 3.17. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1901 - 1950.

Figure 3.18. Comparison of percentage occurrence of SF, LF-N and LF-I for copula vs. semi copula verbs in the time period 1951 - 2000.
Figure 3.19. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1801 - 1850.

Figure 3.20. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1851 - 1900.
Figure 3.21. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1901 - 1950.

Figure 3.22. Comparison of percentage occurrence of SF, LF-N and LF-I for non-reflexive vs. reflexive verbs in the time period 1951 - 2000.
Figures 3.23 to 3.26 show the changes in past tense copula and semi copula verbs with time. They indicate that, in both cases, SF decreases over time whereas LF-I and LF-N both increase. The largest increases are found for LF-I. Changes in the relative occurrence of copula and semi copula verbs follow similar trends. Initially, and across all time periods, there is a significant difference in the proportion (as a percentage) of examples that are of a particular form. In particular, LF-I is more frequent in the semi copula than the copula. This is also the case for the above examples (Figures 3.15 to 3.22), and usually occurs to the relative detriment of the SF. As Figures 3.23 to 3.26 show, however, the trends for both forms, although differing in their proportions, are similar in that the SF decreases while LF-N and LF-I increase. However, the increase in LF is different depending on the verb type. For the past tense copula verb the decrease in the SF results in an increase predominately in the LF-N, whereas the decrease in the semi copula SF results in an increase predominately in the LF-I. Similar trends are apparent in the above examples (copula vs. semi copula, and reflexive vs. non-reflexive, shown in Figures 3.15 to 3.22) they are not as clear as the changes observed in the past tense examples.
Figure 3.24. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1851 - 1900.

Figure 3.25. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1901 - 1950.
Figure 3.26. Comparison of percentage occurrence of SF, LF-N and LF-I for past tense copula vs. past tense semi copula verbs in the time period 1951 - 2000.

Similar trends are not seen for the comparisons of the present tense copula and semi copula verbs (Figures 3.27 to 3.30). The first issue with these data is that the LF-I is absent for the present tense semi copula. So, no clear trends are observed for semi copula examples, with the occurrences across each time period being almost similar, as the decrease in SF, and the resultant increase in LF-N, is relatively small and not significant. For present tense copula examples, the SF decreases significantly and, in the third and fourth time periods, is less prevalent than the LF-I, which increases to the detriment of the SF. Increases in the LF-N are also observed, but they are not as large as those seen for LF-I, and they plateau and stabilize after the second time period.
Figure 3.27. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense zero copula vs. present tense semi copula verbs in the time period 1801 - 1850.

Figure 3.28. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense zero copula vs. present tense semi copula verbs in the time period 1851 - 1900.
Figure 3.29. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense zero copula vs. present tense semi copula verbs in the time period 1901 - 1950.

Figure 3.30. Comparison of percentage occurrence of SF, LF-N and LF-I for present tense zero copula vs. present tense semi copula verbs in the time period 1951 - 2000.
When comparing perfective and imperfective verbs (Figures 3.31 to 3.34), the initial situation (in the first time period examined, 1801 – 1850) is that the predominant forms are SF and LF-1, for both the perfective and imperfective. Changes in the perfective over time result in a decrease in SF which is minimal across the first two time periods, but then drops dramatically, from just over 40% in 1851 – 1900, to 25% in the third time period, and to just over 20% in the final time period. While the LF-N increases slightly over these time periods, the decrease in the SF is mostly at the expense of the LF-I. While a dramatic decrease is also seen over the last two time periods for the imperfective, the trend is not as clear, as SF actually increases from the first to the second time period, and decreases sharply thereafter. As with the perfective, this is mirrored by an increase in LF-I, while LF-N only increases significantly from the first to the second time period, and plateaus thereafter.

Therefore, while the changes in copula and semi copula occurrences can be examined statistically, one of the key outcomes of this study is that, even in a qualitative sense, examination of particular types provides substantially more information on the changes in form over time. Clearly, analysis of this kind will not determine why such forms have a particular tendency to change in the manner that they have, but it does show the clear issues with how the data are analysed, even prior to any interpretation.

Figure 3.31. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1801 - 1850.
Figure 3.32. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1851 - 1900.

Figure 3.33. Comparison of percentage occurrence of SF, LF-N and LF-I for perfective vs. imperfective verbs in the time period 1901 - 1950.
A more detailed statistical examination of the data may be conducted. For example, when comparing the total copula and semi copula verbs, changes in short form may, in this case be compared to those in the combined long form over all four time periods. Alternatively, the long forms only can be compared, and the short form ignored. In addition to $\chi^2$ tests symmetrical and directional measures may be looked at. In essence, a symmetric measure is akin to a correlation coefficient and a value of 0.3 or above suggests a reasonable strength of correlation, but does not carry the positive or negative association of that test, whereas a directional measure shows the strength and significance of the predictive power of each variable examined (i.e. time period or form). One limitation of the symmetric measures tests that follow is that one descriptor, $\phi$, is relevant only for a 2x2 matrix, so it is not relevant to the majority of tests carried out in this particular study. This is shown below, in Figures 3.35 to 3.37. Figure 3.35 shows the output from the SPSS® statistical software, and shows the counts of each form for copula and semi copula verbs. It is consistent in layout to the outputs that yielded the results discussed above.

Figure 3.36 examines the effect of symmetrical measures testing on these data. Firstly, for the data analysed, the term $\phi$ is not, in this case, relevant as its relevance relates to a
2x2 matrix only, whereas this study is predominately on a 2x3 matrix. It shows that the Cramer's V and the Contingency Coefficient are analogous to a correlation coefficient between the two variables being examined. The results of these tests, 0.326 and 0.418, respectively (Figure 3.36) show a good, and significant, correlation between time and form. However, only the magnitude, and not the direction of the change, is returned as a result. Thus, the test does not tell us whether or not a particular form is increasing or decreasing with time, only if it is changing significantly relative to the time period and other forms.

Figure 3.37 shows the output from the directional measures analysis, and the "reduction in misclassification" of data. In essence, this test is telling us what the chances of success are if one was to guess what the form of the adjective was, based on the time period, or the time period based on the type of a verb. For example, according to the Goodman and Kruskal Tau (Figure 3.37), if one was to guess what form a adjective was, the chances of making a mistaken guess would be reduced by 10.2% if one knew what the time period was. Similarly, if one knew the grammatical form of the adjective, one would reduce the chances of guessing the incorrect time period by 7.1%. All these effects are returned as highly significant. However, the percentage reduction in error is roughly equal between time and form for most measures examined.

In terms of further interpretation, the potential uses of directional measures are quite conceptually difficult to interpret. For example, if one considers that the overall frequency of LF-N was 63.2% and LF-I was 36.8%. If one were asked to "guess" what the grammatical form of an adjective chosen at random was, one would be "better off" guessing LF-I, as one would be wrong 36.8% of the time, compared to 63.2% of the time if one "guessed" LF-N. Essentially, the directional measures tell you the reduction in error that would result if one knew the time period. If the directional measure in this study, e.g. Goodman and Kruskal's Tau in Figure 3.37 had a value of 0.102, is considered, this means that the chances of guessing incorrectly would be reduced by 10.2% if the time period was known. In essence, the directional measures provide a way of determining the predictive ability of the two variables, time period or form.
### Time * Form Crosstabulation

<table>
<thead>
<tr>
<th>TIME</th>
<th>COUNT</th>
<th>SHORT.form</th>
<th>LONG form</th>
<th>Long from instrumental</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>19-1</td>
<td>564</td>
<td>50</td>
<td>44</td>
<td>658</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Count</td>
<td>463.7</td>
<td>130.7</td>
<td>63.6</td>
<td>658.0</td>
</tr>
<tr>
<td></td>
<td>% within Time</td>
<td>85.7%</td>
<td>7.6%</td>
<td>6.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% within Form</td>
<td>51.6%</td>
<td>16.2%</td>
<td>29.3%</td>
<td>42.4%</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>36.4%</td>
<td>3.2%</td>
<td>2.8%</td>
<td>42.4%</td>
</tr>
<tr>
<td>19-2</td>
<td>220</td>
<td>84</td>
<td>19</td>
<td>323</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Count</td>
<td>227.6</td>
<td>64.1</td>
<td>31.2</td>
<td>323.0</td>
</tr>
<tr>
<td></td>
<td>% within Time</td>
<td>68.1%</td>
<td>26.0%</td>
<td>5.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% within Form</td>
<td>20.1%</td>
<td>27.3%</td>
<td>12.7%</td>
<td>20.8%</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>14.2%</td>
<td>5.4%</td>
<td>1.2%</td>
<td>20.8%</td>
</tr>
<tr>
<td>20-1</td>
<td>194</td>
<td>122</td>
<td>5</td>
<td>321</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Count</td>
<td>226.2</td>
<td>63.7</td>
<td>31.0</td>
<td>321.0</td>
</tr>
<tr>
<td></td>
<td>% within Time</td>
<td>60.4%</td>
<td>38.0%</td>
<td>1.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% within Form</td>
<td>17.7%</td>
<td>39.6%</td>
<td>3.3%</td>
<td>20.7%</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>12.5%</td>
<td>7.9%</td>
<td>.3%</td>
<td>20.7%</td>
</tr>
<tr>
<td>20-2</td>
<td>115</td>
<td>52</td>
<td>82</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Count</td>
<td>175.5</td>
<td>49.4</td>
<td>24.1</td>
<td>249.0</td>
</tr>
<tr>
<td></td>
<td>% within Time</td>
<td>46.2%</td>
<td>20.9%</td>
<td>32.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% within Form</td>
<td>10.5%</td>
<td>16.9%</td>
<td>54.7%</td>
<td>16.1%</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>7.4%</td>
<td>3.4%</td>
<td>5.3%</td>
<td>16.1%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>1093</td>
<td>308</td>
<td>150</td>
<td>1551.0</td>
</tr>
<tr>
<td></td>
<td>Expected Count</td>
<td>1093.0</td>
<td>308.0</td>
<td>150.0</td>
<td>1551.0</td>
</tr>
<tr>
<td></td>
<td>% within Time</td>
<td>70.5%</td>
<td>19.9%</td>
<td>9.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% within Form</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>70.5%</td>
<td>19.9%</td>
<td>9.7%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

### Chi-Square Tests

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>328.703&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>293.317</td>
<td>6</td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>1551</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.08.

Figure 3.35. Cross tabulation output and results of $\chi^2$ tests from SPSS<sup>®</sup> (v.14) for copula and semi copula verbs.
Nominal by Nominal
N of Valid Cases

Not appropriate for a 2x3 table, only a 2 x 2 matrix

Symmetric Measures

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Approx. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal by</td>
<td>Phi</td>
<td>.460</td>
</tr>
<tr>
<td>Nominal V</td>
<td>.626</td>
<td>.000</td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td>.418</td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>1551</td>
<td></td>
</tr>
</tbody>
</table>

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

These two values are analogous to a correlation coefficient between the two variables

These two values show a reasonable (and highly significant) correlation between time and form

Figure 3.36. Annotated output from the Symmetric Measures test, using SPSS® (v.14). The three boxes highlight the key issues with interpreting this data.

Directional Measures

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Asymp. Sig.</th>
<th>Approx.</th>
<th>Approx. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal by</td>
<td>Lambda</td>
<td>Symmetric</td>
<td>.081</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time Dependent</td>
<td>.123</td>
<td>.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Form Dependent</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Goodman and</td>
<td></td>
<td>Time Dependent</td>
<td>.071</td>
<td>.008</td>
</tr>
<tr>
<td>Kruskal tau</td>
<td></td>
<td>Form Dependent</td>
<td>.102</td>
<td>.016</td>
</tr>
<tr>
<td>Uncertainty Coefficient</td>
<td>Symmetric</td>
<td>.090</td>
<td>.010</td>
<td>6.651</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time Dependent</td>
<td>.072</td>
<td>.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Form Dependent</td>
<td>.119</td>
<td>.013</td>
</tr>
</tbody>
</table>

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Cannot be computed because the asymptotic standard error equals zero.
d. Based on chi-square approximation
e. Likelihood ratio chi-square probability.

Figure 3.37. Annotated output from the Directional Measures test, using SPSS® (v.14).
A similar analysis was carried out by comparing only the long form examples. These results are summarised in Figures 3.38 and 3.40. These data show differences in the LF-I and LF-N over the different time periods. Again, assuming that the null hypothesis defined earlier in this chapter is true, the results show significant differences from the expected ratios. Further, these results indicate that, although both the time period and the form are significantly (i.e. they show a statistically significant change with time) the Goodman and Kruskal Tau and Uncertainty Coefficient (Figure 3.39) indicate that, when form is the dependant variable this results in a stronger power in predicting the LF-I or LF-N. The Tau shows that the chances of making an incorrect prediction of the grammatical form are reduced by 25% if you know the time period. Figure 3.39 shows that the "directionless" measure of correlation shows stronger correlation between time period and form, than for either long versus short or for the entire dataset. These values suggest that the most important correlations – in terms of magnitude – are between time period and LF-N versus LF-I, although all effects are significant.
### Crosstabulation

<table>
<thead>
<tr>
<th>Timeperiod</th>
<th>Count</th>
<th>Nominative</th>
<th>Instrumental</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>19-1</td>
<td>50</td>
<td>50</td>
<td>44</td>
<td>94</td>
</tr>
<tr>
<td>Expected Count</td>
<td>93.2</td>
<td>30.9</td>
<td>94.0</td>
<td></td>
</tr>
<tr>
<td>% within timeperiod</td>
<td>58.2%</td>
<td>46.8%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% within form</td>
<td>16.2%</td>
<td>29.3%</td>
<td>20.5%</td>
<td></td>
</tr>
<tr>
<td>% of Total</td>
<td>10.9%</td>
<td>9.6%</td>
<td>20.5%</td>
<td></td>
</tr>
<tr>
<td>19-2</td>
<td>84</td>
<td>19</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>93.3</td>
<td>33.7</td>
<td>103.0</td>
<td></td>
</tr>
<tr>
<td>% within timeperiod</td>
<td>81.8%</td>
<td>18.4%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% within form</td>
<td>27.3%</td>
<td>12.7%</td>
<td>22.5%</td>
<td></td>
</tr>
<tr>
<td>% of Total</td>
<td>18.3%</td>
<td>4.1%</td>
<td>22.5%</td>
<td></td>
</tr>
<tr>
<td>20-1</td>
<td>122</td>
<td>5</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>95.4</td>
<td>41.6</td>
<td>127.0</td>
<td></td>
</tr>
<tr>
<td>% within timeperiod</td>
<td>98.1%</td>
<td>3.9%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% within form</td>
<td>39.8%</td>
<td>3.3%</td>
<td>27.7%</td>
<td></td>
</tr>
<tr>
<td>% of Total</td>
<td>26.6%</td>
<td>1.1%</td>
<td>27.7%</td>
<td></td>
</tr>
<tr>
<td>20-2</td>
<td>62</td>
<td>92</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>90.1</td>
<td>43.9</td>
<td>134.0</td>
<td></td>
</tr>
<tr>
<td>% within timeperiod</td>
<td>38.8%</td>
<td>61.2%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% within form</td>
<td>18.9%</td>
<td>54.7%</td>
<td>28.9%</td>
<td></td>
</tr>
<tr>
<td>% of Total</td>
<td>11.4%</td>
<td>17.9%</td>
<td>28.9%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>308</td>
<td>150</td>
<td>458</td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>306.0</td>
<td>150.0</td>
<td>456.0</td>
<td></td>
</tr>
<tr>
<td>% within timeperiod</td>
<td>67.2%</td>
<td>32.8%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% within form</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>% of Total</td>
<td>87.2%</td>
<td>32.8%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

#### Chi-Square Tests

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>115.097</td>
<td>3</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>129.726</td>
<td>3</td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>458</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 30.79.

Figure 3.38. Cross-tabulation output from SPSS® (v.14) for analysis of LF-I and LF-N.
Similarly, the combined LF was compared to the SF. Analysis by the $\chi^2$ test indicated a significant difference ($p = 0.000$). The directional measures are different for the Lambda ($\Lambda$) variable, compared to the other tests. The Goodman and Kruskal Tau and Uncertainty Measures both show significant values for time period and grammatical form, with time period being slightly more predictive than form, i.e. when form is the dependent variable. In addition, the symmetric measures show a reasonable correlation between time period and form (Figure 3.40).
This type of analysis, however, does demonstrate the potential of statistical methods to not only characterise the occurrences of predicate adjectives in the Corpus, but also to predict future trends in occurrence. Further, while it may be outside the remit of this particular study, it shows how the dataset may be employed and developed in future studies. The results of the symmetric and directional measures analyses tell us, in effect, that both time period and form show reasonable predictive power for SF versus the LF, but that the time period is highly predictive of the form when nominative and instrumental forms are
analysed (i.e. while all effects of time are significant, there is a stronger association between time period and form when the long form instrumental and nominative data are analysed alone). Also, time period is better at predicting form in this latter dataset than form is at predicting time period.

3.5 DISCUSSION

The application of parametric statistical tests may be limited by the nature of the data being analysed. In particular, it has been observed that the use of parametric tests is valid when the samples are derived from normally distributed populations, where the samples possess similar variances and where the data have been measured in at least an interval scale of measurement. If such conditions cannot be assured, one must seek alternative approaches to analyse the data. This usually involves the use of non-parametric tests, although the use of parametric tests may still be valid in situations where there are only moderate deviations from the above assumptions (for example, this might be where the data may not possess equal variances and the populations from which the data have been sampled may deviate from normality). In addition, the nature of the data measurement is an important determinant in the choice of statistical test. Parametric tests are usually associated with data that can undergo arithmetic manipulations (i.e. a mean value and variance). Therefore, the data must be measured on either an interval or ratio scale. If the data is nominal or ordinal – as is the case in this study – then non-parametric tests should be employed to evaluate compliance, or otherwise, with the null hypothesis.

The $\chi^2$ test is one of a number of non-parametric tests that can be employed to analyse data, given the criteria discussed above. Other tests include the Friedman's test and the Kruskal-Wallis test. The conditions for the use of each of these tests are quite specific, and they are used as alternatives to parametric test methods. For example, when data have been measured on a scale the Kruskal-Wallis and Friedman's tests are the equivalent of the one-way and two-way parametric analysis of variance tests, and are used when the assumptions of ANOVA are not valid. Therefore, where ANOVA has been applied in this study it is with the inference that the data meets the criteria, above, for parametric
distribution of data. Quite often, the label “ANOVA” is used where it actually refers to its non-parametric equivalent, the Friedman's test, which is commonly used for three or more populations. In a similar manner to ANOVA for parametric data, the Kruskal-Wallis and Friedman's tests simply show acceptance or rejection of the null hypothesis and do not show the origins of the suggested differences. In these cases, appropriate post-hoc tests should be carried out. Where such analysis has resulted in unsatisfactory or unclear results, non-parametric tests have been used. In this case, the \( \chi^2 \) test has been used, as the data being analysed can be considered to have been measured as discrete categories within an ordinal scale of measurement. It compares the actual values recorded with those expected according to the null hypothesis. Therefore, in this study the actual frequency of forms returned from searching the Corpus is compared with the null hypothesis, which states that frequency of form will be the same across each pair of time periods compared.

Data collected and analysed by a \( \chi^2 \) test is normally presented as frequencies in discrete categories – in this case, this represents the occurrence of form in each time interval. While a range of tests have been used – from 2 x 2 to 2 x 3 matrices – to analyse data by the \( \chi^2 \) test, the underlying principles of the test are the same. In all cases, the expected number of observations per cell (defined by the null hypothesis, which states that the frequency of form occurrence over time is the same) is compared to the recorded number of observations. From the difference in these cases the \( \chi^2 \)-statistic is calculated, and the significance, or otherwise of a comparison is recorded (i.e. Tables 3.1 and 3.3).

While the \( \chi^2 \) test has many advantages, it is limited by the amount of data required for its correct use – generally, a frequency of five or more is required for analysis. In most cases in this study such a criterion is not an issue, given the size of the dataset employed and the number of examples returned from the Corpus. In a small number of cases, however, the frequency is too low to allow analysis. Consideration therefore is given to how the data has been subdivided, and whether there is any possibility that the data can be grouped differently – for example, it might be appropriate to collate LF-I and LF-N together and compare then directly against the SF. Nevertheless, if such an option is not viable the \( \chi^2 \) test can be carried out with less than the expected frequency. Generally, this produces a test statistic that is larger than might normally be expected when the null hypothesis is true,
increasing the chance of a type 1 error. Conversely, if all the cells with small expected values have an observed frequency very similar to that which is expected they will contribute little to the overall deviance in the results and it is unlikely that the results will be distorted or inaccurate. (Woods et al., 1986: 145). While Woods et al commented that this is something infrequently carried out in linguistics analysis, in all cases where the frequency of samples has been too low to allow a full $\chi^2$ analysis to be carried out this has been described in the relevant tables, or should be apparent from the related figures.

The use of such quantitative statistical methods may be viewed as an attempt to develop previous qualitative and semi-quantitative studies, such as the studies conducted by Benson (1954) and Gustavsson (1976). Such studies were qualitative in design and their results should be viewed in this context. They therefore provide an excellent counterpoint to the stratified and quantitative nature of the current study. For example, Gustavsson (1976) made a synchronic study of predicate adjectives with the copula byt’ in a corpus of 67 texts spanning a period of seven years, encompassing fiction and newspapers sources, in an attempt to determine the rules of adjectival form choice. In this study, he found that, "Sentence or clause type does not appreciably affect choice of form." This would infer that something else instigates the choice of adjectival form. In the present study, this would be verb type. As Gustavsson considered only the copula byt’ in his work, we cannot, therefore, make a direct comparison between the two. For example, he did not consider semi-copula verbs, three examples of which are shown below:

3f.

<table>
<thead>
<tr>
<th></th>
<th>он</th>
<th>остался</th>
<th>непреклонен</th>
<th>и</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>on</td>
<td>osta -l -sja</td>
<td>nepreklon-en</td>
<td>i</td>
</tr>
<tr>
<td>But</td>
<td>he</td>
<td>remain-PST-REFL</td>
<td>inflexible-SF</td>
<td>and</td>
</tr>
<tr>
<td>ero</td>
<td>ne</td>
<td>посмеялся</td>
<td>послушаться</td>
<td></td>
</tr>
<tr>
<td>ego</td>
<td>ne</td>
<td>posme-l -i</td>
<td>osluš -at'-sja</td>
<td></td>
</tr>
<tr>
<td>him</td>
<td>not</td>
<td>dare -PST-PL</td>
<td>disobey-INF-REFL</td>
<td></td>
</tr>
</tbody>
</table>

"But he remained inflexible, and they did not dare to disobey him."

(Bunin)
On next day 3SG.MASC appear-PAST.MASC entirely pale-NOM.MASC

'Next day he appeared completely pale.'

(Turgenev)

The general trends apparent in the data discussed above is that the SF decreases and, in most cases, this is associated with an increase in the LF-I. LF-N also increases across the time periods examined. However, the changes observed for LF-N are not as large as those for LF-I, and they often do not result in specific trends. In most cases, as shown earlier in this chapter, changes in LF-N usually increase from one time period to the next, and thereafter do not change, but plateau at a "steady-state". This is, to a certain degree, a generalisation, and specific changes are discussed in Section 3.4, above. However, in most cases there are statistically significant differences between the frequencies of occurrence of forms across the different time periods.

Figures 3.4 and 3.5. show, respectively, that changes observed in form occurrence over time are more significant for the semi copula examples than for the copula examples. However, if the occurrence of SF is decreasing, it is logical that the occurrence of another form must be increasing. It may be considered that they are due to the aspect of the verb: In essence, it is proposed that the aspect is undertaking the role of temporality associated previously with the SF. Thus, instead of temporality being marked by the SF adjective, it is being marked by the aspect of the verb. Further, because a perfective is used in cases
where "something" is done and finished, or will be done and finished, this will not apply to the present tense. This is in contrast to an imperfective, which is essentially a continuing state or a repeating state. For example, if we consider Figures 3.12 and 3.13, we can see that there is a steady decrease in the occurrence of the SF with the perfective verbs (Figure 3.12) whereas, for imperfective verbs (Figure 3.13) we see what may be described as a "bimodal" distribution – essentially, this means that the first two time periods are reasonably consistent with each other, and so are the last two, but these two pairs are different from each other. This may suggest that the temporality function previously attributed to the SF adjective is now being partly taken over by the perfective aspect.

An important issue to consider is that, with semi copula verbs, LF-I is possible in the present tense, whereas for the copula verbs, where we have a zero occurrence, it is not. Therefore, in the most common predicate constructions there are no examples of the LF-I. This will also possibly skew any statistical analysis, or make analysis difficult due to fewer occurrences. It would therefore seem logical that we may see a greater change within the semi copula verbs as there is a greater scope for change, i.e., the possibility of the LF-I occurring with all forms of the verb. So, if we see a decline in SF and no possibility of this transforming into LF-I usage, does this then result in a concomitant increase in LF-N to compensate for the decrease in SF? If we refer to Figure 3.11, we have a case where the LF-I is not possible. In this case, there is no significant decrease in the SF over time, and nor does the LF-N increase by an appreciable amount. Hence, in this case it may be suggested that, if the grammatical structure does not permit the existence or formation of the LF-I, then the decrease in the SF does not occur. Although this is observed in one set of results in the whole analysis, a certain degree of validation may be drawn from a much wider examination of verb occurrence from the Corpus. Such a qualitative analysis provides us with interesting support for the concept that the decrease in SF occurrence is controlled by the ability of the form to change from SF to, specifically, LF-I, and not to LF-N. This inference is reinforced in Figures 3.31 to 3.34, where the percentage occurrence of form is compared across the time periods for perfective and imperfective verbs. In these examples the trends discussed above are clearly observed. Indeed, such trends are prevalent in all other examples of the data.
The occurrence of LF-N is substantially less than SF or LF-1 across all examples examined. It might be proposed that LF-N is a grammatical form that would appear to be incompatible with semi copula verbs. The nominative case is used for the subject of a sentence. Therefore, it is logical to assume that this would not be the most obvious object for a semi copula verb. Therefore, the trends observed in this chapter might relate only to semi copula, rather than copula, verbs. Extrapolation of the findings of this study outside the “data space” should be conducted with caution, possibly employing both qualitative and quantitative methods of analysis. Thus, in Figures 3.6 and 3.7, we see that reflexive verbs show an increase in LF-1 usage over time, whereas non-reflexive verbs do not show the same trend, remaining reasonably static over the time periods examined. This may be an indication that the reflexive semi copula verbs are following the same trend as verbs such as чувствовать себя (to feel), in that they are taking the instrumental form, and are following a pattern set by other reflexive verbs.

In summary, the results presented in this chapter suggest that the SF is declining significantly, and is being replaced by an increasing occurrence of the LF-1, where the LF-1 is a valid option for the verb to take. Where this is not the case, the SF does not change significantly. The most dramatic changes are observed with semi copula verbs. Such changes may be because these verbs have an overt present tense form with which an LF-1 can be used (clearly, so can the LF-N, but it can be used elsewhere, for example with a zero copula verb) which is not the case with быть. We have also seen a difference between reflexive and non-reflexive semi copula verbs, suggesting that reflexive verbs are more likely to take the LF-1, possibly to draw it into line with other reflexive verbs.
Chapter 4

The influence of nouns in the choice of predicate adjective form.
In Chapter 3 we have examined how the adjective form has changed according to the verb type. In this chapter we consider the differences which occur with subject noun phrases headed by different nouns, namely: concrete, abstract, animate and inanimate. Concrete nouns are things which can be experienced in a sensorial, physical, way. For example, the sky cannot be touched, but it can be seen, hence it is a concrete noun. An abstract noun is something which cannot be experienced sensorially, such as democracy, depression, or love – although love or depression, for example, can be experienced in terms of the emotions relating to either, but they are merely indirect correlates of the named abstraction. For example, love may be manifested through dilation of pupils or physical expressions, but such concrete manifestations are only indirect symptoms rather than something in its entirety. An animate noun is, for example, a person or an animal, specific or general, including proper nouns for names. Inanimate nouns relate to things which physically exist, such as a table or a chair, and may also incorporate abstract nouns, such as those given above. If we consider that the short form is marking a temporary state (e.g. he is cold), it would be logical to assume that the short form would be more prevalent with animate nouns. However, Baženova's work showed that, in scientific texts, the short form is used not only with animate subjects but also when the issue of temporality was not important, for example, "space is infinite" ("Prostrantsvo beskonečno"). This is obviously not a temporary state but something inherent in the nature of space which in other examples would be expressed by the long form nominative.

Further, with regards to animacy and inanimacy, animate subjects will also automatically be concrete. Therefore, an understanding of such matters is important when statistical tests are employed to assess language change. This has been discussed extensively in Chapters 2 and 3. Taken in combination, these techniques indicate the value of using both qualitative and quantitative approaches to examining language change.
In the following pages we have conducted statistical analyses on how the frequency of occurrence of choice of adjectival form changes over time, according to the nature of the noun and how this changes over the two hundred year time period which we are examining.

4.2 AIMS

The aims of this chapter are to analyse, by an appropriate statistical test, or combination of tests, the change in the occurrence of SF, LF-I and LF-N over the two hundred year period which is divided into four parts. This is similar in approach to the work carried out in Chapter 3 except, where Chapter 3 related to verbs, this chapter focuses on the choice of predicate form according to the type of noun phrase.

4.3 METHODS

The methods employed in this chapter are the same as those used, and described, in the previous two chapters. The Corpus was searched for predicate adjectives as before – the same dataset of examples were used as in Chapter 3, except that, in this study, nouns were examined in order to see if there was a correlation between noun type and adjective type. As before, the texts used in this study represent fiction (e.g. novels, short stories), non-fiction (e.g. memoirs, journalism, history) and drama, from the four time periods – 1800 to 1850, 1851 to 1900, 1901 to 1950 and 1951 to 2000.
4.4 RESULTS

Analysis of the Corpus, through examination of the choice of adjectival form with various kinds of nouns (concrete/abstract, animate/inanimate) showed a significant difference. In general, the pattern followed by the nouns investigated was similar to that of the verbs, that is, the SF is increasingly being replaced, in particular by the LF-I.

The majority of examples returned from searches of the Corpus were present in sufficient quantities to allow statistical analysis. The exception to this is the occurrence of the zero copula byt' examples, which were not found in sufficient quantities to allow analysis. The results for occurrences of nouns that could be analysed statistically are presented in Figure 4.1. In general, these results show statistically significant trends in the majority of cases, suggesting that occurrence of form changes over time from SF to LF-I. This is also shown graphically in Figures 4.1 to 4.14. As observed in Chapter 3, results show that the majority of cases where no significant difference is observed predominately occur when the first and second, or the third and fourth, time periods are compared. This suggests that the major change occurred between the second and third periods (spanning 1850-1950), while this rule does not necessarily hold for all the data examined in this study, it does hold in a statistically significant number of cases.

The first group of nouns examined was the combined total of those occurring with the semi copula and copula verbs (Table 4.1). For both concrete and abstract nouns, comparisons of all time periods indicate that the frequency of adjectival form is significantly different in all cases, except where the final two time periods are directly compared (p = 0.088 and 0.795, respectively). Trends for animate and inanimate nouns are not so clear. In the case of the former, comparisons for all time periods are significantly different, except when the first two, and the last two, time periods are compared (p = 0.078 and 0.254, respectively). In the latter case all occurrences of inanimate nouns are significantly different when all time periods are compared, except for the comparison between the second and third time periods (p = 0.527). Interestingly, the same trend, and lack of statistical significance, is also
observed for the occurrence of the concrete inanimate nouns ($p = 0.060$ when the second and third time periods are compared). This would infer that the linguistically significant change in occurrence of forms takes place at the turn of the 20th century. By comparison occurrences of the concrete animate nouns appear to fit the expected pattern observed with most examples, where the significant difference are found in all comparisons except those between the first and second, and the third and fourth, time periods ($p = 0.210$ and $0.275$ for comparisons of 1801 – 1850 with 1851 – 1900 and 1901 – 1950 with 1951 – 2000, respectively).

These results are also shown graphically in Figures 4.1 to 4.6. Figure 4.1 (concrete nouns) shows that the SF decreases and is replaced predominately by LF-I. LF-N increases, but only slightly, and then plateaus. This is the same for abstract nouns, as shown in Figure 4.2, except that the SF is less prevalent at the start (in the first time period, 1801 – 1850) and the LF-I increases more rapidly, reaching a plateau by the third time period. LF-N increases from the first (1801 – 1850) to the second (1851 – 1900) time period but decreases thereafter. Figure 4.3 shows that the frequency of occurrence of animate nouns decreases, but does so in a non-linear manner; the first two time periods (1801 – 1900, combined) show similar occurrences, as do the third and fourth time periods (1901 – 2000), but the latter group shows a significantly lower occurrence of SF than the former. As with the previous examples (Figures 4.1 and 4.2) LF-N increases then decreases again. It may also be suggested that LF-I replaces the SF in most cases. This again reinforces the suggestion that significant linguistic change appears to occur predominately at turn of the 20th century. However the trends are not very clear, as they increase overall after a slight decrease in the second time period (1851–1900). Figure 4.4 shows clear trends for the frequency of occurrence of inanimate nouns, as SF decreases markedly and is replaced mostly by LF-I. LF-N again increases from the first to the second time periods, and then falls away. Similar trends are observed for concrete inanimate examples, shown in Figure 4.5. Trends for change in occurrence of concrete animate examples (Figure 4.6) are not as clear as previous examples, particularly when compared to Figure 4.5 (inanimate nouns). Distribution of the SF is again bimodal, as described above with animate nouns (and shown in Figure 4.3).
Figure 4.1. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula) CONCRETE NOUNS.

Figure 4.2. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula) ABSTRACT NOUNS.
Figure 4.3. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula) ANIMATE NOUNS.

Figure 4.4. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula) INANIMATE NOUNS.
Indeed, Figures 4.5 to 4.8, inclusive, suggest by their varied nature the issues associated with predicting linguistic change, either in the past or in the future, and serve as a representative example of how the implications of this study can be interpreted. These issues are discussed in detail in the following paragraphs. By reference to Chapter 2, and the discussion therein regarding both the probabilistic nature of language change and ability to predict past or future language occurrence based on a particular dataset, these four Figures serve as a reasonable example of how the implications of this study can be interpreted. It also highlights some of the issues and pitfalls in this process. It should be noted that any other Figures from this chapter would hopefully serve the points below just as well.

In Figure 4.5 the trends are clear; for the total copula verbs shown the frequency of occurrence of the SF decreases significantly. Plotting this trend and generating a simple regression analysis suggests that the percentage occurrence of SF drops from approximately 55% (in the 1801 – 1850 time period) to approximately 8% in the final time period examined (1951 – 2000). This trend is linear (with the regression coefficient, $r^2$, being 0.97, an excellent “fit” of the regression line to the data). The trend for change is clearly linear.
Figure 4.5. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; concrete inanimate).

Figure 4.6. Frequency of distribution of SF, LF-N and LF-I across the four time periods for copula verbs (total copula + semi copula; concrete animate).
Figure 4.7. Frequency of distribution of SF, LF-N and LF-I across the four time periods for byf (concrete).

Figure 4.8. Frequency of distribution of SF, LF-N and LF-I across the four time periods for byf (abstract).
Extrapolation of this line outside the range of the data provides two interesting predictions. Firstly, in predicting “backwards” to the preceding time period (1751–1800, which is not examined in this study), such a regression would suggest a 67.3% occurrence of the SF in this time period. It should be noted that such a linear extrapolation will eventually lead to a 100% occurrence of the SF. This is somewhat hypothetical and is dependent on the slope, or gradient, of the line of best fit. Further, as the accuracy of fit diminishes the further one is removed from the data points, error may be introduced into the prediction. The occurrence of LF-I and LF-N have not been modelled or predicted, ostensibly due to the non-linear nature of change observed for LF-N in this example. Change in the SF is measured by extrapolating the straight line backwards until it reaches the time period corresponding to 1751 – 1800, or by solving the equation of the line produced by software packages such as Microsoft Excel for the $y$ term. This seems qualitatively reasonable and sits comfortably with the data collected from the Corpus. However, extrapolating forward to predict future occurrence of the SF is not so simple. In doing this one finds that the regression line will pass through the $x$-axis at approximately 2025. This implies that, for copula verbs, the SF will cease to exist in approximately 15 years. If this is true one simply has to wait to have it confirmed. However, this is highly unlikely in a linguistic sense and serves to highlight the limitations of such modelling on the data. It should also be noted, given the general discussion in Chapter 2, that such predictions are made within suitable bounds of error and most likely occur within a probabilistic distribution rather than having an absolute occurrence.

At this point it is interesting to carry out regression analysis for the last three time periods only: 1851–1900, 1901–1950 and 1951–2000. While a regression analysis with only three points is quite weak and, in some ways, statistically insubstantial, it is an interesting semi-empirical exercise to model these three time periods, and then to try to predict the occurrence of short form in the first time period, which we already know to be approximately 55%. In doing such analysis, we are given the prediction for short form in the 1801 – 1850 time period of 44.7%. Using the same “leave-one-out” technique for the final time period results in a prediction of SF occurrence of 3.7%, again substantially different from the 8% found in the Corpus. These simple, and statistically limited, examples demonstrate how weak such predictions can be, particularly when they are compared to
"real" data. Therefore, any predictions thus made, particularly in the light of extrapolating across so few data points, should be treated with caution. It should also be noted that they relate to only one subset of the data analysed in this study, but nevertheless serve the purpose of illustrating the point. Such trends and predictions might be more accurate if they are modelled using twenty, instead of fifty, year time periods, should sufficient data be available.

A further point of interest is the gradient of the straight line. A gradient, or slope, is determined by dividing the change in the y-axis term (in this case, percentage occurrence of SF) by the change in the x-axis term (year). Consequently, the units of change will be percentage change per year. In this example, the change in occurrence of the SF is approximately 0.3% per year. While the above example predicting SF extinction shows the potential and limitation of such modelling, the determination of rate of change is at least relevant within the range of data analysed and can provide some indication of how dramatic such changes are.

A visual inspection of the SF data presented in Figure 4.5 would suggest that, possibly, an exponential regression is a more appropriate fit — this follows from the appearance of the data, particularly in the second time period, that there might be a possible non-linear trend. Therefore, performing the same analysis as before but replacing linear regression with an exponential term does produce a good line of fit, but one whose $r^2$ value is slightly less, at 0.95, than that found for the linear regression. While the linear regression does provide a better fit to the data such a strong exponential fit is interesting to examine further, from a linguistic point of view.

Compared to the linear regression model, an exponential curve has two interesting features. The first is that extrapolation backwards in time suggests that the occurrence of the SF was, in the time period 1751 – 1800, close to 90%. This is in contrast to the prediction of approximately 67% found when using the linear model. It also shows a flatter progression beyond the final (1951 – 2000) time period and suggests, possibly more realistically, that the SF will not completely disappear if modelled in such a fashion and that it will undergo classical exponential decay, characterised by a long, flat asymptote. Also, in
this case the gradient is not constant and will normally be represented by a differential function (i.e. \( \frac{dP}{dt} \), the change in percentage occurrence over time) which will be different at each point unless the data is re-drawn on a log scale, which will introduce linearity to the data. Again, in the context of the discussion in Chapter 2, this would suggest a more dynamic and less rigid interpretation of how the language changes over time. It is also, conceptually, more in line with a graduated dataset and would lend itself, as does the analysis of sigmoidal language change, to linguistic interpretation in the same manner. Hence, one key point would be to define the beginning and end of the "change event", and where the sampling window (in this study, 1801 – 2000) fits into that. The non-linear modelling is quite controversial, in that it has been cynically commented upon by various authors that anything can be made to fit to a non-linear curve. Indeed, experimenting with Microsoft Excel and the "polynomial" regression function will eventually lead to the production of a model for which a perfect correlation is obtained. While this is clearly both an issue and a facetious and cynical misinterpretation of non-linear modelling, the use of such techniques has value if used in the correct manner. For example, in Rasmussen and William’s work (2006) they clearly advocate the use of the right method for the right set of data. In the case of non-linear models, such as their use of stochastic Gaussian Processes, they underpin the selection of the correct mathematical treatment of their data by the use of associated data modelling techniques, such as principal component analysis (PCA). PCA is a data minimisation technique that can simplify a complex dataset. On one, very rudimentary level, it can determine whether the inherent nature of a dataset is or is not linear. It also identifies the relative significance of each parameter encountered in a complex model. Hence, it can be used as a guide to support the use of linear or non-linear methods as being appropriate for a particular dataset.

Clearly, both models provide useful and interesting interpretations of how the occurrence of SF is changing for this group of criteria. They give no indication of the mechanism of change. However, it is important, as shall be discussed below, to put these comments into context for Figures 4.6 to 4.8, inclusive. It should be noted that the above discussion relates purely to the change associated with the data in Figure 4.5, the distribution of adjectival forms with copula verbs (concrete inanimate nouns) across the four time periods examined. By comparison, the change in the SF of the forms of total copula verbs
(concrete animate) does not show a linear decrease. Rather, a bimodal trend is observed which, if analysed statistically, shows a relatively poor fit to linear regression ($r^2 < 0.8$) and other, various simple methods of analysis. Clearly, therefore, the analysis of concrete inanimate and concrete animate nouns with copula verbs will be different, and this demonstrates differences in the rate of change and in any predictions that may be made from such data. As has been described in Chapter 2, and is further described below, one must also be clear that it is more than possible, given what is known about the history of Slavonic (for example, Larsen, 2005), that the sampling period, as large as it is, may still represent only a fraction of the overall change observed, as Shi (1989) observed in his examination of change in the Chinese language over a one thousand year period. Such an issue will always be relevant when extrapolating outside the range of a dataset.

This may be seen by the use of a simple interactive Gaussian model (http://www.rainsoft.de/projects/gausspro.html). A Gaussian Process may be defined as an infinite dimensional Gaussian Distribution (see Chapter 2) which is normally defined by both a mean and covariance. However, using Gaussian Processes for non-linear regression generally only involves the choice of such a covariance function. Such processes are stochastic and probabilistic in nature and are gaining more widespread use in a variety of fields. In terms of linguistics and the current discussion, such a process provides two useful points. Firstly, the confidence intervals associated with a particular dataset increases significantly the further removed one is from the majority of the data points. This can be modelled on this website for any data a user inputs. Further, the non-linear nature of the Gaussian Process, as well as its use of a probability density function might suggest that it is suitable for analysing complex, multivariate processes in language change, such as the different trends in change associated with different word groupings in this study.

However, further consideration of our examples leads us to Figures 4.7 and 4.8. In these examples we see yet again different trends from those observed in Figures 4.5 and 4.6. In Figure 4.7, the SF decreases slightly but effectively remains around 70 – 80%. Also, LF-I is absent from several of the time periods in Figures 4.7 and 4.8. While this may be an artifact of the paucity of such examples in the Corpus it might also reflect true trends in
occurrence. Without a larger dataset it is not possible to state definitively whether or not this is the case. Such occurrences and absences may skew interpretation and extrapolation of data and should always be considered when interpreting data such as that presented in this study.

Therefore, the four examples chosen highlight a range of positive and negative issues regarding modelling of data, both for the sake of attempting to understand changes and trends, and in the consideration of predictive modelling. Each of the four examples showed distinctly different trends. Another four examples could be chosen and they may show different trends again, or they may show similar trends. Any selection of examples would serve the purpose of this discussion. Collation of data into larger groups may obscure some trends and homogenise the nature of a dataset, or it may have no effect as all the collated examples may exhibit the same propensity for change over time. While a representative sample has, where available, been used, it is clear that the nature of the data available may change, both in volume and overall occurrence, if a different dataset is used, or if the Corpus used for this study is in any way modified or added to. Simplistic modelling by regression analysis should be used with caution, as different methods can give very different results for the same data, and these results can have widely varying interpretations. Nevertheless, underpinning any statistical method with a sound rationale should ensure that any interpretations are clearly made in a very specific context. From the range of trends observed in the data collected in this study, it is clear that individual grammatical groups should be examined in isolation in order to best model change in those groups, unless more complex mathematics may be employed to solve such issues. However, a universal interpretation is also required in order to map, for example, the total decrease of SF across as many examples as possible. Finally, the use of novel methods is recommended for the future as it may allow such complexity to be modelled without losing the individual character of important data subsets, and “machine learning” methods such as the Gaussian Processes described above should allow models to be trained and modified by the inclusion of new data. Such models, including the stochastic Gaussian Processes, underpin the probabilistic nature of the dataset and its inherent fluctuations across the time periods examined.
Therefore, prediction of linguistic change, as discussed in Chapter 2, is a possibility but such predictions need to be modelled and interpreted correctly. Hence, the model should remain as only part of the analysis as qualitative interpretations of change are still clearly required.

Examples of by’t have been shown in Tables 4.1 and 4.2, and also graphically in Figures 4.7 to 4.10. As can be seen from Table 4.2, too few examples of the zero copula were found in the literature to allow a valid statistical analysis to be performed. These results have not been shown graphically. However, an inspection of Table 4.2 shows that, as the LF-I form is one which the by’t zero copula cannot adopt, there are too few examples, particularly of LF-N, to compare with the SF. Also, in common with the findings of Chapter 3, the grammatical unavailability of the LF-I results in little or no change between the SF and the LF-N, suggesting again that, mechanistically, change from SF is not to the LF in general, but to LF-I in particular.

For examples of by’t, the predominant trend is for SF to be replaced by LF-N and not by LF-I, as in most other examples. Figure 4.7 (concrete) shows that the SF decreases, but not significantly. It is replaced by the LF-N, which increases across the first three time periods, and decreases thereafter. Occurrences of LF-I are infrequent, and disappear altogether in the third time period, increasing again in the fourth time period. However, overall occurrence of LF-N and LF-I are rare, in comparison to the occurrence of SF. The same trend is observed for the occurrence of abstract nouns (Figure 4.8), except that even fewer examples of LF-I were found upon searching the Corpus – it is absent completely from the first two time periods and then increases sharply from the third to fourth time periods (from 12.5% to 25%). No real trend is observed for the change in occurrence of the animate (Figure 4.9). In general, SF predominates, oscillating by 5% (from 82.29% to 87.23%) across all four time periods. Changes are compensated for by increases or decreases in the LF, either LF-I or LF-N, as no clear trend is observed among the three forms. As with other examples of by’t, occurrences of LF-I are rarer than occurrences of LF-N. When compared to Figures 4.7 to 4.8, Figure 4.10 (inanimate) provides a significant contrast in that the LF (mostly LF-N) is substantially more prevalent than for concrete, abstract or animate examples. SF decreases across the first three time periods,
equilibrating somewhat in the final time period. LF-N increases from the first to second time period and plateaus thereafter. LF-I is absent from the first two time periods and increases slightly across the third and fourth time periods.

Figure 4.9. Frequency of distribution of SF, LF-N and LF-I across the four time periods for byf (ANIMATE NOUNS)
The small number of samples available may have influenced the analysis of the *by†* (copula and semi copula, Table 4.1) examples. In all cases trends are not clear and this may be due to the small numbers of examples returned from an examination of the Corpus, rather than the result of a particular trend in noun occurrence. In these cases the LF was pooled and compared in a 2x2 matrix, as insufficient examples were found to allow a 2x3 statistical analysis of the Corpus, as was the case with other nouns. These results are perhaps best interpreted qualitatively, or by a combination of quantitative and qualitative methods, and these results are shown in Figures 4.7 to 4.10. It should also be commented that making any specific linguistic conclusions from such results may be somewhat premature in the absence of a greater number of examples to analyse.
<table>
<thead>
<tr>
<th>Type</th>
<th>Time Period</th>
<th>Comparison</th>
<th>( \chi^2 )-squared Result</th>
<th>Cramer's V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.169</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1951 - 2000</td>
<td>0.000</td>
<td>0.304</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.304</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.141</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.411</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.272</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.469</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.209</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.480</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.211</td>
<td></td>
</tr>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.346</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td>Total (semi copula + copula)</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1801 - 1850 vs. 1901 - 1950</td>
<td>0.000</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>0.000</td>
<td>0.527</td>
<td></td>
</tr>
</tbody>
</table>

Note: where too few examples occur, and where it is linguistically sensible to do so, for the \( \chi^2 \) test (n < 5) samples have both been pooled and compared (SF vs. LF-N + LF-I), or the Fisher test performed. This is detailed, where relevant, throughout Section 4.4.
Concrete, abstract, animate and inanimate nouns with a semi copula are also shown in Table 4.1 and in Figures 4.11 to 4.14. They follow the same trends as the previous examples (with the exception of byō) in that significant differences in occurrence are observed across all time periods, with the exception of either the first and second time periods, and the third and fourth time periods. More specifically, occurrence of LF-I is significantly greater for these examples than for those reported above. In addition, while LF-N occurrence is comparatively low compared to the other forms, it is substantially more prevalent for these examples than for those above.

Figure 4.11 (concrete) shows that the occurrence of the concrete sees a decrease in SF and a concomitant increase in LF-I — both trends are bimodal, with change occurring in “blocks” as described above for the first two and final two time periods. This is similar to the trends seen for change in the LF-N which shows a small overall increase over the first two time periods and which plateaus thereafter.

A similar trend is observed for animate examples, shown in Figure 4.13. Figure 4.12 (abstract) shows a similar general trend for LF-I to replace SF, except that the decrease in SF is consistent across the first three time periods, steadying thereafter. While most of the decrease in SF is accounted for by an increase in LF-I, LF-N also exhibits a large increase between the first to second time periods, but decreases thereafter. This is also reflected by a similar change in inanimate examples, shown in Figure 4.14.
Figure 4.11. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (concrete).

Figure 4.12. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (abstract).
Figure 4.13. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (animate).

Figure 4.14. Frequency of distribution of SF, LF-N and LF-I across the four time periods for semi copula (inanimate).
Collectively, the results in Table 4.1 indicate that nouns are, in general, following the same trend in change from SF to LF (and in particular to LF-I) that was seen with verbs in Chapter 3, suggesting that the primary agent of change is the verb rather than the noun. It also shows, in the case of byt’ examples, the issue of interpreting results where a relatively small number of examples are present. In common with the qualitative findings of Chapter 2, it might be proposed that, for quantitative results to be valid, any such analysis will require a "threshold" of occurrences below which statistical analysis is not accurate. In addition, it also indicates the value and importance of a qualitative evaluation of all data so analysed. This is also the case with the examples of byt’ (zero copula) shown in Table 4.2. While a large number of examples exist for SF (particularly for concrete and animate examples), the lack of examples of LF-N – as well as the complete absence of LF-I – ensure that statistical analysis of these data is problematic. While it may be commented that, in the absence of LF-I as a form which the adjective can take, SF does not significantly change (as has been argued in Chapter 3), such a point is substantially stronger when examining verbs due to the number of examples returned from searches. In the Corpus searches for nouns (in this case, examples of zero copula byt’), while the same trend may be observed, the relatively smaller number of examples would suggest that the same conclusions cannot be reached with as much certainty as in Chapter 3.

Table 4.2. Results from corpus searches for types that were returned in insufficient numbers for statistical analysis to be carried out.

<table>
<thead>
<tr>
<th>Type</th>
<th>Time Period</th>
<th>Time</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SF</td>
<td>LF-N</td>
</tr>
<tr>
<td>byt’ zero copula concrete</td>
<td>1801 – 1850</td>
<td>114</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1851 – 1900</td>
<td>132</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1901 – 1950</td>
<td>144</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1951 – 2000</td>
<td>112</td>
<td>6</td>
</tr>
<tr>
<td>byt’ zero copula abstract</td>
<td>1801 – 1850</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1851 – 1900</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1901 – 1950</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1951 – 2000</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>byt’ zero copula animate</td>
<td>1801 – 1850</td>
<td>111</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1851 – 1900</td>
<td>130</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1901 – 1950</td>
<td>135</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1951 – 2000</td>
<td>109</td>
<td>6</td>
</tr>
<tr>
<td>byt’ zero copula inanimate</td>
<td>1801 – 1850</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1851 – 1900</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1901 – 1950</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1951 – 2000</td>
<td>22</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 4.3, below, summarises the statistical results of comparing frequency of occurrence of different types of nouns in the same time period. As in Chapter 3, statistically significant trends are found in particular groups. For example, comparisons of nouns with the copula and semi copula (concrete vs. abstract and animate vs. inanimate) show that significant differences are observed in all cases, except where comparisons are made in the first time period. This would suggest that the frequency of occurrence initially conforms to the null hypothesis (i.e. that frequency of occurrence is the same across all measured pairs) but then changes significantly, suggesting a divergence of form across, in this case, time. This also applies to comparisons in Table 4.3 of concrete animate with concrete inanimate and for semi copula animate and inanimate nouns. The opposite is therefore the case for comparisons of byt' (concrete vs. abstract), where there are significant differences in occurrence in all cases, except for those in the final time period (1951–2000). This would suggest that there is no significant difference in the frequency of occurrence between the adjectival forms in the final time period, compared to the divergence seen in the examples discussed above. In addition, the semi copula (concrete vs. abstract) examples show no significant differences when compared in the first (1801–1850) and last (1951–2000) time periods. This would suggest that the occurrence of these forms were initially the same, diverged and have, in the last time period, converged again.
Table 4.3. Statistical comparison, by $\chi^2$-squared test, of pairs of verbs, in each of the time periods, and the statistical significance of trends.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Time Period</th>
<th>SF</th>
<th>LF-N</th>
<th>LF-I</th>
<th>$\chi^2$-squared</th>
<th>Cramer’s V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1801 - 1850</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (semi copula plus copula): concrete (left) vs. abstract (right)</td>
<td>1801 - 1850</td>
<td>171</td>
<td>85</td>
<td>18 15</td>
<td>75 40</td>
<td>0.379</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>274</td>
<td>83</td>
<td>59 52</td>
<td>67 119</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>117</td>
<td>29</td>
<td>66 26</td>
<td>150 84</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>118</td>
<td>30</td>
<td>63 27</td>
<td>206 139</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1801 - 1850</td>
<td>141</td>
<td>118</td>
<td>11 22</td>
<td>60 65</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>241</td>
<td>116</td>
<td>38 73</td>
<td>89 162</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>77</td>
<td>66</td>
<td>41 51</td>
<td>64 203</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>104</td>
<td>44</td>
<td>40 98</td>
<td>98 246</td>
<td>0.000</td>
</tr>
<tr>
<td>byt’ concrete (left) vs. abstract (right)</td>
<td>1801 - 1850</td>
<td>94</td>
<td>49</td>
<td>10 12</td>
<td>12 0</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>85</td>
<td>27</td>
<td>18 16</td>
<td>4 0</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>41</td>
<td>10</td>
<td>24 11</td>
<td>0 3</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>52</td>
<td>6</td>
<td>19 3</td>
<td>4 3</td>
<td>0.062</td>
</tr>
<tr>
<td>byt’ animate (left) vs. inanimate (right)</td>
<td>1801 - 1850</td>
<td>79</td>
<td>64</td>
<td>5 17</td>
<td>12 0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>75</td>
<td>37</td>
<td>9 25</td>
<td>4 0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>25</td>
<td>26</td>
<td>6 29</td>
<td>0 3</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>52</td>
<td>17</td>
<td>4 18</td>
<td>2 5</td>
<td>0.000</td>
</tr>
<tr>
<td>Concrete inanimate (left) vs. concrete animate (right)</td>
<td>1801 - 1850</td>
<td>49</td>
<td>177</td>
<td>9 19</td>
<td>40 96</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>56</td>
<td>407</td>
<td>33 67</td>
<td>82 174</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>58</td>
<td>129</td>
<td>32 74</td>
<td>176 128</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>17</td>
<td>167</td>
<td>31 76</td>
<td>236 194</td>
<td>0.000</td>
</tr>
<tr>
<td>Semi copula: concrete (left) vs. abstract (right)</td>
<td>1801 - 1850</td>
<td>77</td>
<td>36</td>
<td>8 3</td>
<td>63 40</td>
<td>0.486</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>189</td>
<td>56</td>
<td>41 36</td>
<td>126 119</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>76</td>
<td>19</td>
<td>42 15</td>
<td>150 112</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>66</td>
<td>24</td>
<td>44 24</td>
<td>202 136</td>
<td>0.000</td>
</tr>
<tr>
<td>Semi copula: animate (left) vs. inanimate (right)</td>
<td>1801 - 1850</td>
<td>62</td>
<td>54</td>
<td>6 5</td>
<td>48 65</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>1851 - 1900</td>
<td>166</td>
<td>79</td>
<td>29 48</td>
<td>85 162</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1901 - 1950</td>
<td>52</td>
<td>40</td>
<td>35 22</td>
<td>200 241</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1951 - 2000</td>
<td>63</td>
<td>27</td>
<td>36 32</td>
<td>96 241</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figures 4.15 to 4.18, below, compare concrete and abstract nouns, and show the change in percentage occurrence of form over the four time periods of the study. This indicates that the SF, in both cases, decreases, but the rate of decrease is substantially different, with the concrete showing a decrease in the third time period, whereas occurrences of the abstract lessen in the second time period (1851-1900). While it oscillates slightly, LF-N is generally reasonably stable, increasing in the second time period and remaining reasonably constant thereafter. LF-I appears to replace SF in the majority of cases, and the increase in LF-I mirrors the decrease in SF.
Figure 4.15. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1801-1850.

Figure 4.16. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete vs. abstract nouns (the total of semi copula and copula) in the time period 1851-1900.
Figures 4.17 to 4.22 compare animate and inanimate nouns, and show the change in percentage occurrence of form over the four time periods of the study. As with the concrete and abstract nouns (Figures 4.15 to 4.18), the general trend is for the SF to be replaced by
LF-I. This trend is less prevalent for the animate than the inanimate, where the SF decreases from 59% to 13%. While LF-N increases it does so only slightly, and the change reaches a plateau and varies only slightly, compared to the changes seen for SF and LF-I. Figures 4.21 and 4.22 would appear to indicate that the animate has reached equilibrium and settled into a steady representation of form by the third time period, where the ratio of 4:2:4 (SF:LF-N:LF-I) appears and then recurs in the fourth time period. While the differences in SF and LF-N are not so clear, a similar trend is not seen for the variation of the inanimate over time, which continues to change across all four time periods, with, for example, LF-I increasing from 32% to 72% occurrence from the first to the fourth time periods.

![Graph showing percentage occurrence of SF, LF-N, and LF-I for animate vs. inanimate nouns](image)

**Figure 4.19.** Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1801-1850.
Figure 4.20. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1851-1900.

Figure 4.21. Comparison of percentage occurrence of SF, LF-N and LF-I for animate vs. inanimate nouns (the total of semi copula and copula) in the time period 1901-1950.
There are also marked differences for the change in concrete animate and concrete inanimate nouns over time (shown in Figures 4.23 to 4.26, below). While SF decreases over time for both concrete animate and concrete inanimate, it drops to almost zero for the former, while the decrease in the latter stabilises at approximately 40% from 1901 onwards. Again, similar trends are not observed for LF-N, which initially increases slightly and then decreases again in the fourth time period (1951 – 2000). The majority of the decrease in SF is replaced by an increase in LF-I. However, while the increase for the concrete animate is gradual (from 41% in 1801 – 1850 to 83% in 1951 – 2000) the increase for concrete inanimate is less pronounced (from 33% to 27%, 38% and 44% across the four time periods examined) and is compensated for by a larger increase in LF-N as observed for the concrete animate.
Figure 4.23. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1801-1850.

Figure 4.24. Comparison of percentage occurrence of SF, LF-N and LF-I for concrete animate vs. concrete inanimate nouns in the time period 1851-1900.
Figures 4.25 to 4.34 show the change in occurrence of form of *byt’* (concrete, abstract, animate and inanimate) over the four time periods examined. Bearing in mind the caveats raised above regarding the size of the available examples for analysis, certain trends are clear. For example, Figures 4.27 to 4.30 show the comparative change in concrete and
abstract nouns over time. Occurrence of SF halves across the four time periods for both concrete and abstract, following similar trends. In this case, most of the increase in LF is in the nominative form. However, while LF-N increases up to the third time period, it decreases sharply thereafter. In the case of the abstract, this may be due to the appearance of LF-I – absent from the first two time periods but which grows to 25% of all abstract forms present by the fourth time period, compared to SF (50%) and LF-N (25%) in the final time period (1951 – 2000). LF-I occurrence decreases for the concrete noun over the first three time periods, and increases in the fourth time period. These results should be considered in the context of sample size and total LF occurrence. In the case of the former, as shown in Table 4.2, there are comparatively few examples of byt for zero copula abstract nouns, which may affect the quality of the analysis. More important, however, is the trend shown in Figures 4.27 to 4.30. While this shows a significant increase between the first (1801 – 1850) and second (1851 – 1900) time periods, occurrences of LF plateau thereafter, if LF-I and LF-N are collated. The specific increase in LF-N is clear across the first three time periods, and is characterised by a split in LF occurrence, which results in a decrease, from the third to fourth time periods, in LF-N occurrence and a concomitant increase in LF-I occurrence rather than a recurrence of the SF.

![Figure 4.27. Comparison of percentage occurrence of SF, LF-N and LF-I for byt’ (concrete vs. abstract) nouns in the time period 1801-1850.](image-url)
Figure 4.28. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (concrete vs. abstract) nouns in the time period 1851-1900.

Figure 4.29. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (concrete vs. abstract) nouns in the time period 1901-1950.
Figure 4.30. Comparison of percentage occurrence of SF, LF-N and LF-I for byf' (concrete vs. abstract) nouns in the time period 1951-2000.

Figure 4.31. Comparison of percentage occurrence of SF, LF-N and LF-I for byf' (animate vs. abstract) nouns in the time period 1801-1850.
Figure 4.32. Comparison of percentage occurrence of SF, LF-N and LF-I for 
*byf* (animate vs. abstract) nouns in the time period 1951-2000.

Figure 4.33. Comparison of percentage occurrence of SF, LF-N and LF-I for 
*byf* (animate vs. abstract) nouns in the time period 1951-2000.
Figure 4.34. Comparison of percentage occurrence of SF, LF-N and LF-I for byt' (animate vs. abstract) nouns in the time period 1951-2000.

However, while these trends clearly represent the data returned from the search of the Corpus, the overall low number of data points, compared to other data examined in this chapter, suggests that the results of the change in these nouns should be treated with caution.

This is also the case for the byt' (animate and inanimate) examples shown in Figures 4.31 to 4.34. Clear trends are observed, in that SF declines, being replaced initially by LF-N and, in the final time period (1951 – 2000) by LF-I. However, changes in the animate are often negligible – for example, SF (animate) actually increases in its occurrence from 83% to 87% across the time periods. This example is in contrast with the inanimate, which decreases from 79% to 43% over the same interval, and is replaced initially by LF-N up to the third time period (1901 – 1950) and, increasingly, by LF-I in the final time period, where the occurrence of LF-I rises to 12.5% of the total form count despite its complete absence from the first two time periods examined. Again, such sharp changes and unexpected trends may be a result of the sample size and its potential to skew the data examined.

Krasovitsky and colleagues (2008) commented that the instrumental had already become firmly established by the nineteenth century as a predicate case in copular constructions, appearing consistently under certain structural and semantic conditions, specifically:
Structural conditions:
- Copula in the infinitive
- Copula in the future
- Copula in the non-indicative mood

Lexical semantics:
- Inanimate nouns
- Certain animate nouns which clearly denote temporary states (e.g. свидетель 'witness')

Sentence or contextual semantics:
- Temporal phrase:

(from Krasovitsky et al., 2008).

They further commented that, until the middle of the 20th century, variation in case marking on predicate nouns with the copula byt ‘to be’ was mostly conditioned by semantic and syntactic factors. The second half of the 20th century then saw the spread of the instrumental to domains previously occupied by the nominative, in doing so replacing multiple rules of variation with a single overall rule. This finding is repeated in this study, and is shown in Figures 4.27 to 4.30 (occurrence of SF, LF-N and LF-I for byt (concrete vs. abstract) nouns). In these four figures, the change in occurrence of LF-I and LF-N can be clearly seen to follow the change described by Krasovitsky and colleagues. In Figure 4.29, which represents the period 1901 – 1950, occurrences of the LF are predominately in the nominative. However, in Figure 4.30, representing the period 1951 – 2000, the occurrence of the LF is split between the instrumental and the nominative, in agreement with Krasovitsky’s findings. It may also be the case that the current study also provides examples, such as those discussed above, of where similar rules of variation operate. While changes may occur at different rates in certain lexical classes, such as nouns of nationality, changes may occur more slowly than in the language in general. They concluded that semantics played only a peripheral role in this morphosyntactic process and

154
was ultimately replaced by a single rule, that of instrumental case marking on predicate nouns.

Trends in frequency of occurrence with semi copula examples are shown in Figures 4.35 to 4.42. These result in clearer trends than the examples above. For example, Figures 4.35 to 4.38 compare the change in percentage occurrence of concrete and abstract nouns with a semi copula over the four time periods examined. They show that SF is generally more prevalent for the concrete than for the abstract, and that occurrence decreases (after initially increasing slightly for the concrete) steadily over time, although the decrease in the abstract slows substantially from the third time period (1901 – 1950) onward. In both concrete and abstract occurrence of LF-N increases, across the four time periods, from approximately 5% to 15% in both cases, despite the trends being quite different – whereas concrete occurrence increases steadily and then reaches a plateau from the third time period onwards, abstract occurrence increases sharply from the first to the second time period, and then decreases to the third before increasing again in the final time period. It should be noted that these trends are, overall, relatively small due in comparison to the percentage occurrence of SF and LF-I. The general trend for LF-I is upwards, replacing SF over time, despite a small decrease for the concrete examples in the second time period that mirrors the slight increase seen in SF. This is similar to the trends observed for the change in occurrence of the semi copula animate and inanimate nouns, shown in Figures 4.39 to 4.42. Generally, SF decreases and is replaced by a combination of LF-N and LF-I, but mostly LF-I. However, the magnitude of the trends is substantially less than for other examples. For example, the SF animate decreases only by 11% across the four time periods examined, whereas LF-I increases only by 8%. Further, as Figures 4.39 to 4.42 illustrate, the change is not as a result of a linear increase or decrease in form occurrence, but rather an increase or decrease that effectively splits the four time periods into two distinct parts; in the first two time periods, SF is generally the same; it is also the same in the third and fourth time periods. However, the occurrences in the third and fourth time periods are significantly less than those in the first two time periods. Such a distribution, which is effectively (statistically) bimodal, has been discussed above, and in other chapters, and is clearly the case with the semi copula examples discussed within the timeframe of this study. It is also apparent that changes in LF-N are of a lower magnitude
than those for LF-I, which appears to predominately replace SF, and that changes to LF-N are more variable and follow less clear trends than those for SF and LF-I.

Figure 4.35. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (concrete vs. abstract) with a semi copula in the time period 1801-1850.

Figure 4.36. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (concrete vs. abstract) with a semi copula in the time period 1851 - 1900.
Figure 4.37. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (concrete vs. abstract) with a semi copula in the time period 1901–1950.

Figure 4.38. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (concrete vs. abstract) with a semi copula in the time period 1951–2000.
Figure 4.39. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (animate vs. inanimate) with a semi copula in the time period 1801–1850.

Figure 4.40. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (animate vs. inanimate) with a semi copula in the time period 1851–1900.
Figure 4.41. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (animate vs. inanimate) with a semi copula in the time period 1901–1950.

Figure 4.42. Comparison of percentage occurrence of SF, LF-N and LF-I for nouns (animate vs. inanimate) with a semi copula in the time period 1951–2000.
4.5 DISCUSSION

This chapter has examined the role of different kinds of subject noun in relation to predicate adjectives. The results presented in this chapter can be summarised quite simply. In essence, they show the same general trends in frequency of occurrence as was observed in Chapter 3. This means that the SF decreases over time and is replaced by a combination of LF-N and LF-I. Specifically, we have seen in this chapter that concrete and animate examples are bimodal in their patterns of change. This change is usually characterised by a decrease in SF, which is replaced by an increase in LF, predominately LF-I. By contrast, abstract and inanimate examples show a linear change over all four time periods. This suggests that these findings are not an artifact of the sampling methods or nature of the Corpus, but are linguistically valid, particularly given the size of the sample analysed. Such modes of change may be due to the fact that the type of the noun is not the dominant factor in the choice of the predicate adjective form.

However, in most cases, except those semi-copula examples highlighted in the previous section, SF is replaced by LF-I. Where the change from SF to LF-I is not possible, some change from SF to LF-N is observed. However, change from SF to LF-N is minimal and often inconsistent, generally reaches a plateau by the second time period (1851 - 1900) or does not occur at all. In some cases where change to LF-I is not possible, little or no change from SF to LF-N is observed. This is an interesting result, as SF can freely change to LF-N in other cases — clearly, one must consider why it does not change in these cases, particularly as it is the only option for change. In such cases, the form remains as the SF and does not change to LF-N. The only option other than SF in zero copula examples is LF-N. In such constructions LF-I cannot occur. This is summarised in Table 4.2, where the distribution of form is presented. This table shows that the LF-I is absent, but that the SF does not change to the LF-N, even though it has the option to do so.

It is interesting that, where LF-N is the only option for change, the vast majority of forms remain as SF. This may be because the verb cannot take nominative objects. This is
acceptable for SF as it is not a regular nominative, but rather a highly specialised form of
the nominative. This is interesting in the context of findings by Krasovitsky and colleagues
(2008), suggesting that the spread of the instrumental that they described, and the
associated replacement of multiple rules with a single overall rule, may not be possible, or
be limited, in cases where the nominative is the only long form that the adjective can adopt.
Change to LF-I is not possible for zero copula (i.e. present tense semi copulas, discussed
above); if there was an overt form of the verb byt’ in the present tense in Russian one can
assume that we would be seeing the change towards LF-I est’. This is in contrast to other
literature that shows the SF decrease being mapped by an increase in LF-N. This has been
discussed extensively in Chapter 1, and may relate to the construction and analysis of
different copora. For example, this study has examined the Barentsen Corpus, which
contains approximately 31,000,000 words. In Chapter 1, the corpora used were smaller in
size, often by a number of orders of magnitude compared to the Barentsen Corpus. They
were also subjectively compiled – while this may also be said of the Barentsen Corpus, this
might be offset by the size of the Corpus, whereas in smaller corpora studies the size and
scale of the examples available to the researcher may inadvertently skew the data and any
conclusions drawn from it. However, as discussed in Chapter 1, it should be made clear
that the Barentsen Corpus was assembled independently of this project without input from
researchers in this project. Therefore, the Corpus should be free from the bias, intentional
or otherwise, of researchers on this project. It should also be noted that this corpus was
compiled in order to investigate changes in the Russian language over the period 1800 –
2000 but not specifically morphosyntactic change.

Baženova, for example, examined scientific texts only. Scientific papers are written in an
overly formal manner, often incorporating examples of subject-specific meta-language
which may be grammatically eccentric in that it is comprehensible to a small audience but
whose meaning, particularly the use and context of certain words, may not be clear to a
wider readership. Examples of such language are discussed in Section 2.2.2. The use of a
resource as large and varied in input sources as the Barentsen Corpus removes any such
issues from analyses carried out upon it. In such cases, compiling a corpus from such texts
alone may not result in an analysis of form that is consistent with other sources, even
though they may clearly illustrate differences in language use, as shown in Section 2.2.2.
This argument may, for example, relate to computer terminology, where the quality and nature of the output is dependent upon the quality and nature of the data input, or on which the study is based. Similarly, if a corpus is constructed from one type of text, conclusions into wider, more holistic language change, cannot be made or validated. This is a widely applicable principle common to many fields and, again, may define the limitations present in some previous corpus studies. On the other hand, collation of all types of data into one large corpus may itself present an issue. Specifically, inconsistencies and seemingly different trends may be due to the different effects, including weightings, within a corpus which may skew the results. So, smaller studies can be important, whether quantitative or qualitative, as they can focus on a specific area, whereas a large corpus study, such as this one, can return generalised trends. However, it should be noted that taking representative samples from a larger dataset can mean that some of the more interesting individual examples may not be found purely because they have fallen outside the sample set. Hence, careful sampling of a suitable population is the key to consistent and valid quantitative or qualitative results.

The above point on sample size relates to the description and discussion of form change observed in Figures 4.7 to 4.10. While these results are linguistically interesting, and possibly counter-intuitive in terms of adoption of LF-N and not LF-I, the nature of the samples examined is such that no definitive conclusion can be drawn. This may be a definite trend in language change, or it may be an artifact of availability. In these examples, relating to byt’, LF-N is seen to increase substantially more than LF-I. This, as briefly described above in Section 4.4, may be a realistic reflection of change but it may also be due to a relatively small set of examples being returned from a search of the Corpus. These examples were statistically analysed by combining the LF-N and LF-I together, as there were insufficient examples, particularly of LF-I, to conduct the same statistical analysis as with the other examples.

Bimodality of the change in occurrence of form is often observed. Again, one must consider the points raised above in terms of the veracity of such change, as it may be an artifact of the sample, or a specific linguistic event. However, examination of Figures 4.1 to 4.4, and Figures 4.11 to 4.14 show the same trends are apparent. In essence, and in a qualitative
sense, concrete and animate examples are bimodal in their patterns of change (usually a
decrease in SF, which is replaced by an increase in, predominately, LF-1) whereas abstract
and inanimate examples show linear change over all four time periods. This suggests that
these findings are not an artifact of the sampling methods or nature of the Corpus, but are
linguistically valid. This is supported by the inter-relationship of the pairs described above.
This is an important point in the validation of statistical methods for analysis of the Corpus.

Clearly, there is a different rate, and style, of language change for these examples. Such
trends are also observed in Chapter 3, but are not as consistent or clear as those in this
chapter. However, the perceived bimodality may be due to the rate of language change,
and such descriptions are qualitative and relate to empirical observations of the data. In the
case of the predicate adjective with concrete plus abstract examples, the rate of change
may be occurring more slowly than for those that show a linear change in frequency of
occurrence of different forms. In essence, this suggests that the four time periods are,
effectively, only two periods (19th century versus 20th century) and that a true comparison
with the other examples might require analysis over a longer time period. This is illustrated
in Figure 4.43, below, which compares schematic examples of bimodal and linear
decreases in occurrence of form, and suggests that the bimodal trend may be indicative of
the sampling period. In effect, while the reasons for defining clear strata into which the time
period of this study (1801–2000) has been divided, it must be remembered that the data –
and trends – may not automatically fit such compartments. It should also be noted that the
changes observed in this study might represent a small fragment of a larger change
occurring over a longer period of time, as discussed above in the context of the Slavonic
languages. Thus, any trends described might only be a small part of a larger trend and may
or not be representative of that larger phenomena – it is simply impossible to determine
this without conducting a larger study.

With this in mind, therefore, Figure 4.43, below, shows several ways in which the data may
vary across the time period of this study. It does not represent how the data may change
beyond the boundaries of this study, although a certain amount of extrapolation, both pre-
1801 and post-2000, is statistically viable but may well result in an increased variance and
reduced confidence in the accuracy of any predictions made outside the time period of this
study. The trendline drawn through each of the bar charts represent different ways in which
the occurrence of forms have been shown to change in this study. Part (a) shows a linear
decrease, where the gradient would allow a rate of change (with the units of percentage
occurrence per year, for example) to be estimated. Part (b) shows a bimodal distribution of
the data. Such a non-linear change is more difficult to model and use for predictive
purposes. Part (c) shows a hybrid of (a) and (b) where a certain degree of bimodality is
combined with a general decrease in occurrence. This demonstrates a possible limitation
of fitting trendlines to such data. It should also be noted that the x-axis in part (c) is based
on a hypothetical 400-year sample, whereas (a) and (b) are based on 200-year samples.
Such consideration is important when discussing the implications of studies of this nature,
particularly in the extrapolation and contextualisation of results outside the range
examined.

Therefore, in conclusion, clear trends in occurrence of forms have been observed. The SF
decreases and is replaced by, in most cases, the LF-I. Where the LF-N is the only option
for change the form in most cases does not alter, although a slight increase in LF-N has
been inconsistently observed. Occurrence of form – either as a decrease or an increase –
appears, in a qualitative sense, to be either linear or bimodal, although the latter may be an
artifact of the sampling period. This may be due to the fact that the type of the noun may
not necessarily be the dominant factor in the choice of predicate adjective form.
Figure 4.43. Schematic representation of proposed types of language change. Part (a) shows a typical linear decrease in occurrence, whereas part (b) shows the bimodal decrease. Part (c) shows how, if examined over a longer time period, the bimodal trend may in fact be part of a larger linear decrease in occurrence.
Chapter 5

The influence of lexical item in the choice of predicate adjective form.
5.1 INTRODUCTION

In work carried out on predicate nouns within the Short Term Morphosyntactic Change project, Krasovitsky et al (2008) used a corpus-based approach to investigate the change in predicate nouns in Russian. They found that semantics played only a subsidiary role in the morphosyntactic process. It was found to only support variation at some stages, particularly in the first half of the 19th century. Ultimately, it was being replaced in favour of a single rule which saw instrumental case marking on predicate nouns. The shape of the nominative–instrumental shift indicated that Russian is moving from a semantically-conditioned to a syntactically-determined model for predicate nouns, and that this change is at an advanced stage in the contemporary language. This has been discussed in Chapter 4, particularly where similar trends in the adoption of the instrumental were observed.

In the previous chapters of this thesis, we have examined predicate adjectives with regard to verbs and subject nouns. It has been observed, and discussed in Chapter One, that there is a group of adjectives which favour the short form in the predicate position:

болен  bolen sick
виноват vinovat guilty
готов gotov ready
dalek  dalek far
dovolen dovolen satisfied
dолжен dolžen necessary/must
dостоин dostoin worthy
здоров* zdorov healthy
намерен nameren intend
нужен nužen necessary
похож pokož similar
прав  prav correct
свободен svoboden free
склонен sklonen inclined
согласен soglasen agreeable
способен sposoben capable
счастлив* sčastliv happy
(in the above list, an asterisk denotes adjectives added for the purpose of this study to lists given in previous works. All adjectives are given here in the masculine short form.)

In addition, as the adjectives zdorov and sčastliv have been added as a result of empirical observation of texts in the Corpus, rather than listed in previous works and dealt with in a systematic or consistent manner, we might expect these particular adjectives to have a more flexible predicate form. These examples might need to be analysed separately, or they may skew a larger dataset if they are present in a large enough number. Alternatively, they could be analysed in a simple, qualitative manner.

5.2 AIMS

The adjectives listed in Section 5.1 are those that commonly take the SF. Therefore, the aim of this study is to determine if those adjectives have increased in their use of SF, to the detriment of LF-N and LF-I. Therefore, results of searches from the Corpus were analysed, by an appropriate statistical test, or combination of tests, in order to assess any significant change in the occurrence adjectival forms (SF, LF-N and LF-I) over the two hundred year period which is divided into four parts. This is similar in approach to the word carried out in Chapters 3 and 4 except, where Chapter 3 related to verbs and Chapter 4 to subject nouns, this chapter focuses on the choice of predicate form according to its adjective class.

5.3 METHODS

The methods employed in this chapter are the same as those used, and described, in the previous two chapters. The Corpus was searched for predicate adjectives as before – the same dataset of examples was used as in Chapter 3 in order to see if there was a correlation between the trends in noun type and adjective type. Analysis was carried out by collating all adjectives found in the search, as well as individually analysing changes in
each adjective separately. As before, the texts used in this study represent fiction (e.g. novels, short stories), non-fiction (e.g. memoirs, journalism, history) and drama, from the four time periods – 1800 to 1850, 1851 to 1900, 1901 to 1950 and 1951 to 2000, including the list in 5.1 above.

5.4 RESULTS

5.4.1 Preliminary studies and findings from a qualitative analysis of adjective class and predicate form

It is believed that the predicate adjective class may be splitting, in that certain lexical items, such as gotov, zdorov and so on, may now be favouring the short form in the predicate position, where other adjectives still require a certain element of choice. Rather than concentrate on individual lexical items, it was thought necessary to see if this split was occurring along the lines of the adjective classes described in Dixon and Aikenvhald (2004), and the classes are given as follows:

1. Dimension (big, small)
2. Age (old, new, young)
3. Value (good, bad, important)
4. Colour
5. Physical Property (hard, strong, sick, tired)
6. Human Propensity (jealous, happy, clever)
7. Speed
8. Difficulty
9. Similarity
10. Qualification
11. Quantification
12. Position
13. Cardinal Numbers, and “first” and “last”.

169
The above classes are listed in each of the pie charts on the following pages. Dixon identifies the first four, Dimension, Age, Value and Colour, as being the “four core semantic types” (p. 3), which can be found in most languages. Physical Property, Human Propensity and Speed are described as “peripheral”, but are still substantial groups in their own right. The remaining six are listed as “being associated with large adjective classes in some languages” (p. 5).

In order to test the hypothesis, a search was made of predicate adjectives in drama texts. All predicate adjectives found were then extracted and collated into Table 5.1, below.

<table>
<thead>
<tr>
<th>Example SF LF IF Adjective Class</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>быстрый (belyj)</td>
<td>1</td>
</tr>
<tr>
<td>бедна (bedna)</td>
<td>1</td>
</tr>
<tr>
<td>белые (belye)</td>
<td>1</td>
</tr>
<tr>
<td>бесцветный (bescvetnyy)</td>
<td>1</td>
</tr>
<tr>
<td>бойкое (bojkoe)</td>
<td>1</td>
</tr>
<tr>
<td>бедное (blednoe)</td>
<td>1</td>
</tr>
<tr>
<td>богатые (bogatye)</td>
<td>1</td>
</tr>
</tbody>
</table>

Once the items had been grouped according to the classes 1-13 above, a count was made, firstly of the total number of predicate adjectives and, secondly, the numbers in each of the classes. These were further divided according to time period and form of the adjective.

Figure 5.1, below, shows the overall comparison of predicate adjectives for the four time periods. The SF is the favoured form, showing only a small decline in the period 1901-1950, but dropping sharply (by over 20%) in the second half of the twentieth century. LF-I usage remains fairly constant, but LF-N usage shows a sharp increase in the final quarter of the two centuries. This result is in line with previous studies which have suggested the decline of the SF, in particular with regard to its replacement by the LF-N.
Below, in Figures 5.2 to 5.9, the results of dividing the adjectives according to the classes defined by Dixon are shown in more detail. First, the relative proportions for each of the classes in predicate usage were examined. There are two pie charts for each period. The first shows the proportion for each adjective class according to the total number of tokens used (for example, if molod 'young' has been used 3 times in the SF, it is counted as three items in class 2). The second shows the proportion based on the individual lexical items used (in this case, molod, even if used three times, will only be counted once in class 2). In calculating the proportions in this way, we should see if adjective class usage is skewed by the use of particular lexical items which belong to those classes.
Figures 5.2 and 5.3, above, show the results from Griboedov. A small difference occurs between the two charts in that the proportion of class 6 adjectives has decreased (from 45% to 41%), showing that more of the items belonging to that class are being used more than once. For class 5, however, the proportion has increased from 12% to 20%,
suggesting that fewer of its members are being used more than once. There is little—if any—change in the proportions for the other classes.

Figure 5.4. Relative proportions (%) of predicate adjectives according to class for Turgenev, representing the block 1851–1900.

For the second half of the nineteenth century there is something of a reversal, as the proportion of class 6 adjectives increases (from 40% to 44%) and those of class 5 decrease (from 19% to 15%) (Figures 5.4 and 5.5). Once again, the proportions for the
remaining classes remain reasonably stable, although each of the smallest (classes 1, 2, 8, 9, 10) all increase slightly, by 1% each.

Figure 5.6. Relative proportions (%) of predicate adjectives according to class for Gor'kij, representing the block 1901-1950.

Figure 5.7. Relative proportion (%) of lexical items according to adjective class for Gor'kij, representing the block 1901-1950.
By the beginning of the twentieth century, Figures 5.6 and 5.7, we see that the classes 3, 5 and 6 still dominate in terms of numbers of predicate adjectives, but the differences between figures 19 and 20 are barely noticeable – in fact, the proportion of adjectives in class 6 does not change, being 32% in each case. From this, we can conclude that there is a greater number of lexical items being used as predicate adjectives, and that these are not being repeated as frequently as those in the previous two time periods (1801-1850 and 1851-1900).

![Figure 5.8](image1.png)

Figure 5.8. Relative proportions (%) of predicate adjectives according to class for Petrusevskaja, representing the block 1951-2000.

![Figure 5.9](image2.png)

Figure 5.9. Relative proportion (%) of lexical items according to adjective class for Petruševskaja, representing the block 1951-2000.
Finally, in the second half of the twentieth century, Figures 5.8 and 5.9, we see that — once again — there are changes in classes 5 (from 44% to 33%) and 6 (from 32% to 40%), but these occur alongside a change in class 3 adjectives more significant than in the previous three time periods (from 10% to 16% in the second half of the twentieth century).

In conclusion, adjective classes 3, 5 and 6 show the most change regarding the number of lexical items in each class as measured against the frequency of the usage of each item, while the other adjectives types show little or no change. However, it should be remembered that classes 5 (Physical Property) and 6 (Human Propensity) in particular represent a large number of potential members — certainly more than, for example, class 2 (Age) and class 7 (Speed).

Figure 5.10 to 5.13, below, represents the percentage usage of SF, LF-N and LF-I in the predicate, in each of the adjective classes, over the four time periods. As with the previous pie charts, each of the percentages calculated are taken from the total number of predicate adjectives, not from the total number of lexical items in the texts used.

Figure 5.10. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Griboedov, representing the block 1800-1850.
Figure 5.10 shows the dominance of the SF in the first half of the nineteenth century. The majority of predicate adjectives take the SF in this period, with few appearing in the LF-I and no examples for LF-N. The dominant adjective classes are 3 (Value) and 6 (Human Propensity), representing 37% and 54% of the total number of predicate adjectives, respectively. There are no examples of predicate adjectives in classes 1 (Dimension), 4 (Colour), 8 (Difficulty), 10 (Qualification) and 13 (Cardinal Numbers).

![Number of predicate adjectives (%)](image)

Figure 5.11. Distribution of SF, LF-N and LF-I for each adjective class. Taken from Turgenev, representing the block 1851-1900.

In the second half of the nineteenth century, shown in Figure 5.11, classes 3 and 6 still dominate, but class 5 adjectives now account for 20% of predicate adjectives. The LF-I appears only in class 6, and the LF-N appears in class 3. The SF remains the dominant predicate adjective form. However, there is little substantial change regarding the predicate adjective in this period as compared to 1801–1850.
By the first half of the twentieth century, shown in Figure 5.12, a substantial change has occurred. The SF is no longer the dominant form in any of the adjective classes, and has been overtaken by the LF-N. Predicate adjectives appear in more classes, with only classes 7 (Speed) and 13 (Cardinal Numbers) lacking any examples. The LF-N is now the most popular form, appearing in 9 of the 11 classes containing predicate adjectives. The LF-I is also becoming more popular, appearing in classes 1, 2, 3, 5 and 6, accounting for an equal percentage with the SF in class 2 (1%). The SF appears most frequently of the three forms in class 3 (18%) and has equal status with the LF-N in class 6 (21%). Such trends continue to evolve, and are shown for the second half of the 20th century in Figure 5.13.
The above graphs give some evidence to suggest that adjective class is an important consideration in the choice of the predicate form. However, it must be noted that the data were taken from a small group of drama texts, and that the same examination of fiction and non-fiction texts would give different results. For example, the LF-I does not feature particularly strongly in any of the results, but we would expect to find a far greater number of them within non-fiction texts, due to their more "bookish" style. Therefore, the results not only provide some indication of the importance of adjective classes, but also show the need to compare results across genres and to look at those results as a whole.

5.4.2 Qualitative analysis of the Corpus: Predicate adjectives with copula and semi-copula verbs

In this section the nature of the data returned from searching the Corpus is explored. Prior to developing or considering quantitative statistical methods of analysis, the nature of the
data is examined. This relates as much to the prevalence of form as to the way in which it is examined, and how conclusions are drawn. In investigating predicate adjectives with copula and semi-copula verbs, it was found that, while there were plenty of examples with the copula byt', and many with the semi-copulas stat' and kazat'sja, many other semi-copulas returned very few results – certainly not enough with which to carry out a statistical analysis. Therefore, the results from the copula byt' have been divided along genre lines, and the semi-copulas have not. This has been done in order to give as much uniformity to the presentation of the results as possible.

Therefore, the aim of this section is to track any changes which have occurred in predicate adjective usage with copula and semi-copula verbs over the period 1800 – 2000, and to consider the particular elements which may have affected this change. This section will also be used to show how corpus data can be analysed, and the potential issues associated with either qualitative or quantitative analysis.

The data examined below relates to occurrences of twenty verbs returned from searches of the Corpus, following the process described above. These occurrences have been collated and are presented, as percentage of total occurrence, in Figures 5.18 to 5.37. Tables 5.2 to 5.4 present the counts of each item returned from the searches of the Corpus, and Table 5.5 summarises some simple regression tests carried out on the data, in order to establish whether or not any significant trends were observed.
Table 5.2. Summary of occurrence of SF for each verb in the Corpus.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Verb translation</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>бывать</td>
<td>to be, be present</td>
<td>45</td>
</tr>
<tr>
<td>делаться</td>
<td>to become, get</td>
<td>8</td>
</tr>
<tr>
<td>сделяться</td>
<td>to become, get, grow</td>
<td>5</td>
</tr>
<tr>
<td>казаться</td>
<td>to seem, appear (to be)</td>
<td>24</td>
</tr>
<tr>
<td>показаться</td>
<td>to appear</td>
<td>1</td>
</tr>
<tr>
<td>показываться</td>
<td>to appear</td>
<td>0</td>
</tr>
<tr>
<td>оказываться</td>
<td>to prove (to be)</td>
<td>0</td>
</tr>
<tr>
<td>оказаться</td>
<td>to prove (to be)</td>
<td>3</td>
</tr>
<tr>
<td>оставаться</td>
<td>to remain, stay</td>
<td>8</td>
</tr>
<tr>
<td>оставаться</td>
<td>to remain, stay</td>
<td>9</td>
</tr>
<tr>
<td>получаться</td>
<td>to prove, to turn out (to be)</td>
<td>0</td>
</tr>
<tr>
<td>получить</td>
<td>to prove, to turn out (to be)</td>
<td>0</td>
</tr>
<tr>
<td>представляться</td>
<td>to occur, seem</td>
<td>0</td>
</tr>
<tr>
<td>представляться</td>
<td>to occur, seem</td>
<td>0</td>
</tr>
<tr>
<td>прибыть</td>
<td>to arrive</td>
<td>0</td>
</tr>
<tr>
<td>прибыть</td>
<td>to arrive</td>
<td>0</td>
</tr>
<tr>
<td>становиться</td>
<td>to become, grow</td>
<td>15</td>
</tr>
<tr>
<td>стать</td>
<td>to become</td>
<td>2</td>
</tr>
<tr>
<td>явиться</td>
<td>to appear</td>
<td>0</td>
</tr>
<tr>
<td>явиться</td>
<td>to appear</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5.3. Summary of occurrence of LF-N for each verb in the Corpus.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Verb translation</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>бывать byvat'</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>делаться delat'sja</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>сделаться sdelat'sja</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>казаться kazat'sja</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>показаться pokazat'sja</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>показываться pokazyvat'sja</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>оказываться okazyvat'sja</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>оказаться okazat'sja</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>оставаться ostanovit'sja</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>оставаться ostanovit'sja</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>получаться poluchit'sja</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>получиться poluchit'sja</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>представляться predstavit'sja</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>представляться predstavit'sja</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>прибывать pribyvat'</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>прибыть pribyt'</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>становиться stanovit'sja</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>стать stat'</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>являться javlit'sja</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>явиться javlit'sja</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5.4. Summary of occurrence of LF-I for each verb in the Corpus.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Verb translation</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>бывать</td>
<td>to be, be present</td>
<td>1</td>
</tr>
<tr>
<td>делаться</td>
<td>to become, get</td>
<td>6</td>
</tr>
<tr>
<td>сделать</td>
<td>to become, get, grow</td>
<td>6</td>
</tr>
<tr>
<td>казаться</td>
<td>to seem, appear (to be)</td>
<td>9</td>
</tr>
<tr>
<td>показаться</td>
<td>to appear</td>
<td>7</td>
</tr>
<tr>
<td>показываться</td>
<td>to appear</td>
<td>0</td>
</tr>
<tr>
<td>оказываются</td>
<td>to prove (to be)</td>
<td>1</td>
</tr>
<tr>
<td>оказаться</td>
<td>to prove (to be)</td>
<td>10</td>
</tr>
<tr>
<td>оставаться</td>
<td>to remain, stay</td>
<td>18</td>
</tr>
<tr>
<td>остаться</td>
<td>to remain, stay</td>
<td>6</td>
</tr>
<tr>
<td>получаться</td>
<td>to prove, to turn out (to be)</td>
<td>0</td>
</tr>
<tr>
<td>получиться</td>
<td>to prove, to turn out (to be)</td>
<td>0</td>
</tr>
<tr>
<td>представляться</td>
<td>to occur, seem</td>
<td>1</td>
</tr>
<tr>
<td>представиться</td>
<td>to occur, seem</td>
<td>0</td>
</tr>
<tr>
<td>прибывать</td>
<td>to arrive</td>
<td>0</td>
</tr>
<tr>
<td>прибыть</td>
<td>to arrive</td>
<td>0</td>
</tr>
<tr>
<td>становиться</td>
<td>to become, grow</td>
<td>9</td>
</tr>
<tr>
<td>стать</td>
<td>to become</td>
<td>2</td>
</tr>
<tr>
<td>являться</td>
<td>to appear</td>
<td>1</td>
</tr>
<tr>
<td>явиться</td>
<td>to appear</td>
<td>1</td>
</tr>
</tbody>
</table>
Figures 5.14 and 5.15, above, show the general trends for all twenty verbs examined in this particular study. Figure 5.14 shows the trends for SF, LF-N and LF-I, whereas Figure 5.15 combines LF-N and LF-I to compare the trends between SF and total LF. Clearly, in the latter case, as percentage change is being examined, there is an inverse relationship between the SF and LF verb. Figure 5.14 provides more information, decoupling LF-N and LF-I, and shows the differences in appearance, over the time period examined, of these
verbs. In Figure 5.14, it is clear that appearance of LF-I is more pronounced than LF-N. Therefore, the overall trend across the time period is to see a decrease in the usage of SF verb, which mirrors a concomitant increase in LF, particularly LF-I, usage.

The data was examined in two ways. Firstly, the occurrence of ten verbs (see Tables 5.2, 5.3 and 5.4, above) was found by examining the entire corpus. Secondly, the occurrence of a different ten verbs was found by sampling the Corpus randomly. These results were combined to show general trends (Figures 5.14 and 5.15). Both of these subsets of the data were examined individually, and the results are shown in Figures 5.16 and 5.17.

![Figure 5.16](image1.png)

**Figure 5.16.** Changes in SF, LF-N and LF-I usage for verbs where the whole corpus was searched (a total of ten different verbs).

![Figure 5.17](image2.png)

**Figure 5.17.** Changes in SF, LF-N and LF-I usage for verbs where a subset of the Corpus was searched (a total of ten different verbs, mutually exclusive from those shown in Figure 5.16).
The trends apparent in Figure 5.16 are consistent with those observed in Figure 5.14. In this particular case, LF-N and LF-I occurrence are more different than for the overall dataset (i.e. compare the relative positioning of the red and green lines in Figures 5.14 and 5.16) but the same general trends are again observed. However, while Figure 5.17 generally shows comparable trends, the occurrence of LF-N does not significantly change throughout the time period examined. Therefore, both sets of results (Figures 5.16 and 5.17) demonstrate that occurrence of SF decreases significantly across the time periods examined, with a particularly significant transition taking place between the 1851 – 1900 and 1901 – 1950 time periods. Essentially, the 1801 – 1850 and 1851 – 1900 time periods are statistically similar, as are the 1901 – 1950 and 1951 – 2000 time periods. These two groups are significantly different from each other. Such trends are more readily apparent for SF and LF-I, whereas occurrences of LF-N appear to change less over the time periods examined. They also suggest that the exact changes seen may be dependent upon the occurrence of particular verbs in the dataset. In such circumstances, the nature of the sample taken for analysis is of considerable importance.

In practical terms, it is impossible to examine every single example of an adjective in the Corpus. This is impractical and time-consuming. Therefore, appropriate methods of sampling, which at their core are based on random selection of examples, are vitally important in developing a robust quantitative or semi-quantitative approach to corpus analysis. Therefore, anecdotal usage of the Corpus is inappropriate. For example, if a researcher is aware that a particular verb has a particular occurrence, and the researcher decides to include this information in the “random” sample, then the sample is potentially skewed. Such practice may compromise the methodology used, where the selection of a representative and random sample is key to the successful analysis of the data. In information technology, the phrase “garbage-in, garbage-out” links the poor quality of data input to the model with poor results and potentially misleading conclusions. Therefore, validation of the methodology is vital before any linguistic conclusions can be drawn from this corpus study.
For the results shown in Figures 5.16 and 5.17, a combination of LF-N and LF-I, plotted against SF, simply gives a mirror-image of the trend shown by SF. Therefore, these trends are not shown beyond Figure 5.14.

Regression analysis of the trends obtained for SF and LF-I show clear trends in the data across the four time periods examined (Table 5.5, below). However, when estimating the lines of best fit, it is not clear if changes observed best fit linear or exponential models, as the correlation coefficients are similar, at around 0.95 in all cases. For LF-N, the changes appear to fit best to a logarithmic model; in such cases, the $r^2$ value is 0.95, compared to 0.86 for the next best fit (linear).

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Adjective Type</th>
<th>Short Form</th>
<th>Long Form Nominal</th>
<th>Long Form Instrumental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Nominative</td>
<td>0.84</td>
<td>0.80</td>
<td>0.55</td>
</tr>
<tr>
<td>Linear</td>
<td>Instrumental</td>
<td>0.85</td>
<td>0.87</td>
<td>0.56</td>
</tr>
<tr>
<td>Logarithmic</td>
<td></td>
<td>0.81</td>
<td>0.89</td>
<td>0.51</td>
</tr>
<tr>
<td>2nd Order Polynomial</td>
<td></td>
<td>0.85</td>
<td>0.88</td>
<td>0.56</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td>0.70</td>
<td>0.87</td>
<td>0.47</td>
</tr>
</tbody>
</table>

In general, the results shown in Figures 5.14 to 5.17 clearly show step-wise changes in the occurrence of SF and LF adjective. Specifically, SF occurrence decreases significantly after 1900, whereas LF, and in particular LF-I, shows a significant increase across the same time period and at the same point as changes are observed in SF. Changes in the specific verbs examined are shown, and discussed, below.
Figure 5.18 Percentage change in occurrence with *быть* "to be, be present" (509 total examples; 49, 234, 87 and 139 in each time period).

Figure 5.19. Percentage change in occurrence with *делаться* "to become, get" (124 total examples; 15, 36, 16 and 87 in each time period).
Figures 5.18, 5.19 and 5.20 are best viewed initially as a set, as a significant contrast can be made between their trends based purely on the size of the available data subsets. These are listed in the legends for each Figure. When a sufficient number of verb examples are present, clear trends are obtained. Figure 5.18 is constructed from a total of 509 examples which were found after searching the whole corpus. Similarly, Figure 5.19 shows a similar general trend to Figure 5.18, although the trend for LF-N is not as distinct as in Figure 5.18. This, however, could be considered to map to the generalized trends shown in Figures 5.14 to 5.17. Figure 5.20, on the other hand, is constructed from 21 examples. No clear trends are observed, and one whole time period (1801–1850) did not have any examples for SF, LF-N or LF-I, despite the whole corpus being searched. The lack of available data would therefore make any comments about changes with this particular verb difficult to substantiate. Therefore, this data would suggest that obtaining a minimal number of examples to allow a full and representative analysis is vital.

All three of these Figures were constructed after a full search of the entire corpus. Clearly, the available data can significantly affect the results obtained, and the trends extrapolated.
In essence, the composition of the Corpus will decide the results obtained. They would also suggest, in a wholly qualitative sense, that a minimum number of examples is required before a consistent, reproducible trend can be ascertained.

Figure 5.21 is similar to Figure 5.20, in that a small sample size (14 examples) yields data, and subsequent trends, that are incomplete and unrepresentative of the overall dataset. This would suggest that the sample size of these verbs is too small in the Corpus to allow individual analysis. However, given the overall trend for the whole dataset, it might be suggested that, were more examples present in the Corpus, then results and trends in occurrence similar to those observed in Figures 5.18 and 5.19 might begin to emerge. It should also be noted that Figure 5.20 is comprised of data from the whole corpus, whereas the data in Figure 5.21 is from a sample of the Corpus. Both have similar sample sizes, and both are equally unrepresentative of the overall trends observed. Therefore, as will be seen subsequently, sample size, and taking a representative sample for analysis, is central to producing a meaningful and valid analysis.

Figure 5.21. Percentage change in occurrence with представиться представить'sja 'to occur, seem'
(14 total examples; 1, 7, 3 and 3 in each time period).
However, the results in Figure 5.22 would suggest that the generalizations discussed above are not wholly representative. In Figure 5.22, the verb was found 426 times after a partial search of the Corpus. A more detailed examination of Tables 5.2, 5.3 and 5.4 suggest that the data for this verb is not evenly distributed. For example, in the second time period (1851 – 1900) there are only 19 examples, compared with 214 in the next time period (1901 – 1950). This would suggest that, in order to obtain representative data for individual verbs, it is important to have as even a distribution of data as possible, as well as a minimum threshold of data available.
Figure 5.23. Percentage change in occurrence for оказаться okazat'sja 'to prove (to be)' (57 total examples; 13, 23, 7 and 14 in each time period).

Figure 5.24. Percentage change in occurrence for оказываться okazyvat'sja 'to prove (to be)' (77 total examples; 1, 45, 9 and 22 in each time period).
Again, Figures 5.23, 5.24 and 5.25 illustrate the absence of both evenly distributed data, and a sufficient volume of data. Both verbs have less than 100 examples (including that shown in Figure 5.23, which followed a search of the whole corpus), and some time periods where few, or no, examples, were found. Consequently, no clear trends are discernible. This is also independent of whether or not the whole corpus was searched; for example, Figure 5.23 is comprised of a search of a fragment of the Corpus, whereas Figures 5.24 and 5.25 are constructed from data obtained after full searches of the Corpus.
Figures 5.26 and 5.27 are therefore notable for having a reasonable distribution of data and in having in excess of 100 examples of each verb. They also show trends that, while not entirely in agreement with the overall picture of verb change, are broadly suggestive of
such trends. Compared to Figures 5.23, 5.24 and 5.25, for example, Figures 5.26 and 5.27 provide a significant improvement in the distribution of data. Again, this is suggestive of the importance of the size of the dataset in achieving a realistic understanding of changes in verb occurrence throughout the Corpus.

Figure 5.28. Percentage change in occurrence for получиться получиться ‘to prove, turn out (to be)’ (92 total examples; 0, 2, 15 and 75 in each time period).

Figure 5.28 also exhibits an uneven distribution of data, making it difficult to discern any clear trends for adjective occurrences in this case. Figure 5.29 shows a similar lack of discernible trend with a similarly sized data set (92 for получиться polučit'sja, and 99 for показаться pokazat'sja).
Представляться *представляться* (Figure 5.29, and Tables 5.2, 5.3 and 5.4) is an example of where the distribution of data is as important as the sample size. While a total of 142 examples, derived from a search of the whole corpus, would appear to be sufficient, 92 of these (approximately 65% of the dataset for this verb) relate to examples of LF-I in the time period 1851 – 1900 only. This lack of evenly distributed data also makes the estimation of any clear trends in the data difficult. In clear contrast, Figure 5.30 - *сделать* - shows clear trends in verb occurrence. This may be due to the sample size – 217 – and the reasonably even distribution, compared to examples above, of data across the four time periods – 51, 68, 71 and 27 examples, respectively. This example also shows a clear difference in the changing occurrence of LF-N and LF-I across the time period examined. Similar results are also obtained for *становиться* (shown in Figure 5.31. The only issue with *становиться* is the relatively low number of examples of LF-N, although clear trends for SF and LF-I are apparent. Figure 5.32, by way of contrast, shows again similar issues to earlier examples, where the sample size and distribution appear to be insufficient across the whole time period being examined.
Figure 5.30. Percentage change in occurrence for сделать́сь *sдела́ться* 'to become, get, grow' (217 total examples; 51, 68, 71 and 27 in each time period).

Figure 5.31. Percentage change in occurrence for станови́ться *становиться* 'to become, grow' (294 total examples; 26, 79, 34 and 155 in each time period).
Verbs показываться pokazyvat'sja, явиться javit'sja and приывать pribyvat' were present in such low numbers that graphs are not shown for them. Tables 5.3, 5.4 and 5.5 summarise the occurrences of these verbs. In addition, while получаться poluchatsja had a reasonable number of examples, their distribution was such that no clear trends could be established. Again, the entire corpus was searched for this verb, and reference should be made to Tables 5.3, 5.4 and 5.5 for the distribution of these data.

Therefore, in summary, clear qualitative trends in the changing occurrence of SF, LF-I and LF-N verbs are discernible across the twenty verbs examined. Collectively, clear trends are apparent, with decreases in SF being matched by increases in LF, notably LF-I, while LF-N increases by a relatively smaller amount, and in some cases decreases. Changes in individual verbs are more difficult to discern due to the frequency and distribution of examples in the Corpus, and the absence of examples from particular time periods may skew the data in a manner that is not reflected when the data for all verbs is collated. However, where these factors are adequate (generally, a dataset subset for each verb of at least 150 – 200 example, evenly distributed across the four time periods of the study), trends in verb occurrence may be predicted.
5.4.3 Initial Search of the Corpus – Analysis of Collated Adjectives

The aim of this particular study was to determine if the usage of common SF adjectives, listed above in Section 5.1, has increased over the last two hundred years. Therefore, a search of the Corpus was carried out for each of these adjectives in the manner described earlier in this thesis. The output of examples was processed as before and collated into a single dataset and analysed statistically ($\chi^2$-test, $p = 0.05$) by the methods described in Chapters 3 and 4. The results are given below in Table 5.6, and presented graphically in Figure 5.33.

Table 5.6. Summary of statistical tests for predicate adjectives where the complete dataset was analysed by collating all the selected adjectives together.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Form</th>
<th>Comparison</th>
<th>$\chi^2$-squared</th>
<th>Cramer's V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SF</td>
<td>LF-N</td>
<td>LF-I</td>
<td>Result</td>
</tr>
<tr>
<td>1801 - 1850</td>
<td>96</td>
<td>2</td>
<td>5</td>
<td>1801 - 1850 vs. 1851 - 1900</td>
</tr>
<tr>
<td>1851 - 1900</td>
<td>725</td>
<td>76</td>
<td>31</td>
<td>1801 - 1850 vs. 1901 - 1950</td>
</tr>
<tr>
<td>1901 - 1950</td>
<td>235</td>
<td>24</td>
<td>2</td>
<td>1801 - 1850 vs. 1951 - 2000</td>
</tr>
<tr>
<td>1951 - 2000</td>
<td>316</td>
<td>27</td>
<td>15</td>
<td>1851 - 1900 vs. 1901 - 1950</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1851 - 1900 vs. 1951 - 2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1901 - 1950 vs. 1951 - 2000</td>
</tr>
</tbody>
</table>
Figure 5.33. Percentage change in adjective occurrence for SF, LF-N and LF-1 over time, based on the data presented in Table 5.6.

Figure 5.33, above, presents the change in adjective form over time (as a percentage), and shows that there is virtually no significant change in occurrence across the four time periods examined. Table 5.6 does, however indicate that statistically significant differences were observed across most time periods (where \( p < 0.05 \) in Table 5.6) except when both the first and fourth, and the second and fourth time periods were examined. Generally, this would indicate that there is little consistent trend to change for these examples as we have seen for nouns and verbs in previous chapters. While this may be an effect of the number of samples returned from the search, particularly in the occurrence of LF-1 in two cases and LF-N in one, it is, given the overall size of the dataset examined, likely to be due to a linguistic issue, and not a sampling issue or a statistical anomaly. For example, if particular adjectives favour the short form – the list of adjectives examined in this study are those that have been shown to favour the SF, and the data presented in Table 5.6 would therefore suggest that this is the case – and if these adjectives are fossilising into the SF then it should be expected that the percentage occurrence of the short form would increase or remain reasonably constant, and this would be mirrored by a concomitant decline, or stasis, in LF-N and LF-1 usages. However, given the nature of the statistical significance of...
the results, as described above, the usage of SF does appear to decrease significantly in some cases.

There are advantages and disadvantages to separating all adjectives examined in this study and analysing them individually. However, it was discussed in Section 4.6 that any detailed information may be skewed or invalidated by the number of occurrences returned from the search. In Table 5.6 it is apparent that this would certainly be an issue for the long form if the total count, as the data presented in Table 5.6, subdivided for each of the twenty adjectives, returns very few examples for the long form (both nominative and instrumental) and in some cases returns no examples at all. Therefore, evaluation of such small subsets may be limited qualitatively (as discussed in Chapter 2) or quantitatively (as discussed in Chapters 3 and 4). Hence, further examination may simply result in information that provides incomplete or inaccurate information on the change of form over time, and should be treated cautiously as it may result in inferences being drawn that cannot be fully substantiated. Such an analysis is summarised in Table 5.7, and illustrated in Figures 5.34 to 5.50, where the change in occurrence of each adjective analysed in this study is presented.
Table 5.7. Summary of predicate adjective occurrence in the Corpus.

<table>
<thead>
<tr>
<th>Adjective</th>
<th>Time</th>
<th>SF</th>
<th>LF-N</th>
<th>LF-I</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonly used and</td>
<td>1801-1850</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.185</td>
</tr>
<tr>
<td>Single-word form</td>
<td>1851-1900</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.389</td>
</tr>
<tr>
<td>Long-form</td>
<td>1901-1950</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.201</td>
</tr>
<tr>
<td>Multiplicative form</td>
<td>1951-2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
</tr>
<tr>
<td>Total</td>
<td>1801-1850</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.185</td>
</tr>
<tr>
<td>Commonly used and</td>
<td>1801-1850</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.389</td>
</tr>
<tr>
<td>Single-word form</td>
<td>1851-1900</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.201</td>
</tr>
<tr>
<td>Long-form</td>
<td>1901-1950</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
</tr>
<tr>
<td>Multiplicative form</td>
<td>1951-2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
</tr>
<tr>
<td>Total</td>
<td>1801-1850</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.389</td>
</tr>
<tr>
<td>Commonly used and</td>
<td>1801-1850</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.389</td>
</tr>
<tr>
<td>Single-word form</td>
<td>1851-1900</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.201</td>
</tr>
<tr>
<td>Long-form</td>
<td>1901-1950</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
</tr>
<tr>
<td>Multiplicative form</td>
<td>1951-2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
</tr>
<tr>
<td>Total</td>
<td>1801-1850</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.389</td>
</tr>
</tbody>
</table>

Note: The table data shows the frequency and statistical significance of predicate adjectives in different time periods.
5.4.4 Analysis of change for individual adjectives

Figure 5.34 shows the changes over time in the occurrence of болн болен 'sick'. The main change observed is that the occurrence of SF increases significantly between the first and second time periods (1801 – 1850 and 1851 – 1900), and remains relatively high – and significantly different – compared to the first time period. A concomitant drop in LF-N and LF-I is observed across the four time periods, although LF-N decreases in the second time period (1851 – 1900) and increases thereafter, from approximately 5% of all occurrences to approximately 18%, almost reaching the level observed in the first time period (20%). It should be noted that the sample size returned from the search of the Corpus is, for LF-N and LF-I, relatively small, and any results should be interpreted quantitatively with this limitation in mind.

Figure 5.34. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective болен болен 'sick'.

Figure 5.35 shows the changes over time in occurrence of виноват vinovat 'guilty'. Again, while quantitative interpretation might be hampered by the shortage of examples, particularly of the LF-N, it is clear that the majority of the examples returned from the
search of the Corpus are for the SF, which is responsible for between 88 and 100% of all occurrences across the four time periods examined in this study. While LF-N and LF-I therefore occur only rarely, it should be noted that, after no examples were returned from a search of the first time period (1851 – 1900), LF-N and LF-I then contribute to a total of approximately 12% of the all occurrences in the second time period (1851 – 1900). While their occurrence decreases thereafter, it should be noted that such a trend may be an artifact of the Corpus, in terms of its content or size, or it may be a true reflection of the change in occurrence of vinovat. Table 5.6 also indicates that the occurrences of forms are not significantly different between any of the time periods.

Figure 5.35. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective виноват vinovat 'guilty'.

This is similar in some way to the trends in the occurrence of готов gotov 'ready', shown in Figure 5.36. SF is responsible for most occurrences and while this dips slightly in the second and third time periods (1851 – 1900 and 1901 – 1950) it increases again in the final time period (1951 – 2000), being replaced predominately by LF-N. Again, a wholly quantitative interpretation of these results may be limited due to the small number of samples returned from the search of the Corpus.
A slightly different trend is observed with далеко dalek 'far, distant', shown in Figure 5.37. While SF again predominates, in this case by increasing to a plateau by the second time period (1851 – 1900), the occurrence of LF-N is, at approximately 20%, substantial in the first time period. It decreases thereafter, often appearing in the second and third time periods with LF-I, which has the lowest occurrence. LF-I is also absent from the first and fourth time periods examined in this study. By contrast, occurrences of довolen dovolen 'satisfied, content', shown in Figure 5.38, show an opposite trend, in that SF decreases slightly in the final time period from a plateau and is replaced by an increase in LF, particularly LF-N which is the only long form present in the final two time periods. Such differing trends are unlikely to be due to the volume of examples returned from the search (250 and 307 total examples for dalek and dovolen, respectively) and therefore may be an artifact of the small samples returned for LF-N and LF-I, or to a particular linguistic phenomena.
Figure 5.37. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective далек 'far, distant'.

Figure 5.38. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective доволен 'content, satisfied'.
Another wholly different trend is shown in Figure 5.39, where должен должен 'should' is shown to be present predominately as the SF. Either long form is present rarely, if at all, being comprised of approximately 0.8% of all occurrences of должен. While this trend may not be due to a lack of examples – the search of the Corpus having returned 1607 examples in total, it may be related to their distribution. For example, there are very few occurrences of LF-N and LF-I, as is clear from an inspection of Table 5.6 or Figure 5.39, and approximately half of all examples relate to examples of the SF from just one time period (1851 – 1900). Indeed, Figure 5.10 (occurrences of намерен намерен 'intend') is beset with similar issues. In this case, however, the complete absence of any examples of LF-N or LF-I ensures that, other than reporting the relative occurrences of SF, LF-N and LF-I, any further analysis is impossible.

![Figure 5.39](image-url)

**Figure 5.39.** Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective должен должен 'necessary'.

A very different trend is observed for changes in the occurrence of достоин достоеин 'worthy, deserving', shown in Figure 5.40. SF appears to be less dominant than it is for the other adjectives examined, compared to LF-N and, in particular, LF-I, which is present in
three time periods at up to approximately 20%. Overall, SF increases from approximately 60% of occurrences to 80% in the third time period, and decreases thereafter. The long form is represented in all time periods, and mostly by the instrumental form, which is more prevalent than the nominative form, where it is present at all (i.e. in all time periods except the third, 1901 – 1950). However, no significant differences in frequency of occurrence are found from the $\chi^2$ test (Table 5.6). This may be due to the overall small sample size – 56 – which was returned from the search and analysed. Interestingly, a similar trend is observed for occurrences of здоров здоров ‘healthy’, in Figure 5.41. A similar trend to that observed for достой достой is shown. In the case of здоров здоров the return of a total of 250 examples from the search of the Corpus, compared to 56 for достой достой may suggest that the statistical interpretation of the observed trends is more substantive than for достой достой. Similar trends are observed for здоров здоров as for достой достой. However, in the latter case examples of LF-N and LF-1 are more prevalent, and LF-N is, in all four time periods, greater than or equal to LF-1 in its occurrence. As is shown in Table 5.6, there are no significant differences observed in the frequency of occurrence across all four time periods.

![Figure 5.40](image)

**Figure 5.40.** Occurrence, as percentages, of SF, LF-N and LF-1 across the four time periods of this study for the adjective достой достой достой ‘worthy, deserving’.
A similar trend is also found for нужен nuzen 'necessary', shown in Figure 5.43. In this case, 1094 examples were returned from the search of the Corpus, and show a similar trend in occurrence across the time periods to dostoin and zdrov. In the case of nuzen LF-
N is, in three of the four time periods, more prevalent than the LF-I. In addition, the frequency of occurrence in the first time period (1801 – 1850) is significantly different to all other time periods. This can be clearly seen in Figure 5.43, particularly in the decrease in LF-I but, more prominently, in the decrease in LF-N, particularly between 1801 – 1850 and 1851 – 1900. In comparing the trends for dostoin (Figure 5.40), zdorov (Figure 5.41) and nuzen (Figure 5.43), while they are broadly – in a qualitative sense – similar, it is interesting to observe how this is reflected in their statistical analysis (Table 5.6). When the qualitative analysis of such data was described in Chapter 2, sample size was considered to be an important issue, particularly in the emergence of clear, representative trends from what otherwise might be considered to be the background noise of a small and unrepresentative sample. In the three examples compared above, similar trends are observed in frequency of usage across the four time periods, and this begins to show statistical significance when the sample size is at its largest. This may, therefore, indicate the importance of sample size in such analyses, and highlights the limitations of examining each of the adjectives examined in this study individually, as opposed to collectively (shown in Figure 5.33). Therefore, it can be argued that what is developing is a fossilization of the form, in that these particular lexical items have shown a definite tendency to adopt the SF – a process which has precedents in the Slavonic languages, as both Polish and Czech have a small number of short form adjectives for specific lexical items. In this instance, the vast majority of adjectives in Polish and Czech no longer have the possibility of taking a SF, but maintain some frequently used lexical items in this form (interestingly, the word for 'glad' takes a SF across most Slavonic languages).
Figure 5.43. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective нужен \textit{nužen} 'necessary'.

It is also, therefore, interesting that analysis of a reasonably large sample of 673 occurrences of похож \textit{poxož} 'similar', shown in Figure 5.44, returns a very different result to those described above. In the case of \textit{poxož}, LF-N and LF-I are far more dominant than for most of the other adjectives examined. In some cases, such as the middle two time periods, the total occurrence of both long forms is greater than the SF, something that only occurs rarely in the rest of this particular dataset (i.e. the first time period (1801 – 1850) for the adjective способен \textit{sposoben} 'talented, capable', shown in Figure 5.17). Indeed, the only significant differences in occurrence are found between the second time period and the third (1851 – 1900 and 1901 – 1950) and between the second time period and the fourth (1851 – 1900 and 1951 – 2000). This is interesting because the distribution shown in the fourth time period for \textit{poxož} is, generally, that which dominates the rest of the dataset. This might suggest that \textit{poxož} behaves differently to the other adjectives examined in this study, or it may suggest that its rate of change, or that the timing of its change is not the same as the other adjectives examined. It might also be observed that \textit{poxož} appears to "settle", or reaches a point of equilibrium in its tendency to change, in the final time period into the same frequency of occurrence as the other adjectives. It finally demonstrates, for
example, the type of distribution across forms exhibited by прав prav ‘right, correct’ and согласен soglasen ‘agreeable’, shown in Figures 5.45 and 5.47, respectively, across all four time periods. In the case of prav, for example, the relatively small sample size – particularly for LF-N and LF-1 – makes any further analysis difficult. It also shows an opposite trend to свободен svoboden ‘free’, shown in Figure 5.46. In this case, the relative stability of the first three time periods is replaced with a greater divergence of form, with the SF contributing approximately 50% of occurrences in this time period. While the $\chi^2$ test suggests that only the third and fourth time periods exhibit significantly different distributions, this may be due to the gradual but small decrease in both LF-N and LF-1 from the first time period to the third, as well as the complete absence of LF-1 from the first time period. It may also, again, reflect the total sample size (167) returned from the analysis of the Corpus and the veracity of any conclusions thus drawn. Such an issue is also relevant to склонен sklonen ‘inclined’, shown in Figure 5.47. The total sample size of thirteen makes any analysis of frequency of occurrence, or any other phenomena, impossible. Similarly, no clear trend is observed for the change in occurrence of способен sposoben ‘capable’, shown in Figure 5.49. Although the sample size is larger than for sklonen (112 compared to 13) no clear trends can be discerned, with all forms showing increases and decreases throughout the four time periods examples. While such trends may represent non-linear change in usage, they do not fit the overall trends observed for adjective occurrence and, without a larger sample, it is difficult to substantiate such a claim without a larger dataset. If the trends for sposoben are compared to счастлив счастлив ‘happy’, shown in Figure 5.50, it may be suggested that the larger sample size is responsible for apparently removing any skew from the data. Счастлив shows similar trends to many of the examples above (Figures 5.34 to 5.49) and is in broad agreement with the general trend, which suggests that the SF remains the dominant form across all four time periods, changing only for specific adjectives, whereas occurrences of LF-N and LF-1 are comparatively minor and, similarly, show no specific trend across the time periods.
Figure 5.44. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective "похож poxo2 'similar'."

Figure 5.45. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective "прав prv 'right, correct'."
Figure 5.46. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective свободен svoboden 'free'.

Figure 5.47. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective склонен sklonen 'inclined'.
Figure 5.48. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective согласен *soglasen* 'agreeable'.

Figure 5.49. Occurrence, as percentages, of SF, LF-N and LF-I across the four time periods of this study for the adjective способен * sposoben* 'capable'.

215
We have considered the details of Figures 5.44 to 5.50 and seen that the SF is still strong with these lexical items, although a small decrease can be seen in some (sčastliv, sposoben) while others barely change throughout the 200 years (soglasen, prav). This may be attributed to the ability of the adjective to stand in the attributive position – happy and capable can both be used attributively:

| способный | студент          |
| sposobn-yj | student        |
| capable-NOM | student |
| счастливый | ребёнок       |
| sčastliv-yj | reběnok        |
| happy-NOM | child          |

However, the LF-N of prav has a different meaning, and would, in the attributive position, generally mean "right" as in "right-hand side", whereas the SF means "right" as in "correct". The logical position for the SF's meaning is with a copula verb, as something or someone has to be correct.
The first striking feature of the analysis presented in Table 5.7, particularly if compared to Tables 3.1 and 4.1, is the large number of examples that exist in such small numbers that a full statistical analysis is not possible. In the case of verbs, Table 3.2 was constructed to show the examples that could not be analysed due to small number of occurrences in the Corpus. The treatment of adjectives in this chapter is slightly different. This is because the adjectives being chosen are those that have been shown to prefer the SF, and as such the occurrences of LF-N and LF-I should be expected to be less, particularly if the SF is thought to increase in usage to the detriment of LF-N and LF-I. Table 5.7 also includes the results from the Fisher Exact Test, generated automatically by SPSS® software and which is used in cases where too few examples exist for analysis by the $\chi^2$ test.

With regard to the data itself, Table 5.7 and Figures 5.33 to 5.50 show some consistent trends, particularly upon inspection in a qualitative sense. Compared to the results – and the trends observed – in Chapters 3 and 4, such substantial trends are not observed in respect of changes to adjectival form. In most cases the general trend is that, summarised across all time periods, SF is the dominant form (accounting for 82.22% of all examples of adjectives returned from searches of the Corpus), with LF-N (11.40% of all examples) and LF-I (6.38% of all examples) contributing smaller proportions of the total number of examples. Therefore, to generalise, most examples returned from a search of the Corpus show the same trend over the four time periods used in this study. The general trends are, in most cases, statistically similar, indicating a lack of any real trends over time. This is perhaps not surprising given the predominant role of the SF in such adjectives. However, given the size of the overall sample size and compared to the verbs and nouns explored in Chapters 3 and 4, the absence of any real trend, whether it be migration to the SF or to either LF, is surprising. Of course, the caveat to the above is the overall sample size examined, and the lack of examples across parts of the dataset, particularly for LF-N and LF-I.

Therefore, the change in adjective usage observed in this study is minimal over the time period examined. This is consistent with expectations and differs substantially from the trends in verb and noun form occurrence over the same time period.
5.4.5 Stress and stress alternation in the SF

In order to check the effects of stress alternation (that is to say that the LF-N and SF have different stresses; there is no implication that one changes to or from the other) on adjective choice – or, more accurately, to see if there was any correlation in written Russian – a sample of 402 individual adjectives the final quarter (1951–2000) was selected, and the stress patterns checked using Borunova’s pronunciation dictionary (Borunova et al, 1983), as it is a specific stress-pattern dictionary. This was done by identifying all possible combinations of stress patterns for the SF, giving the 16 variations listed below:

<table>
<thead>
<tr>
<th></th>
<th>m+</th>
<th>f+</th>
<th>n+</th>
<th>p+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m+</td>
<td>f+</td>
<td>n+</td>
<td>p+</td>
</tr>
<tr>
<td>2</td>
<td>m+</td>
<td>f+</td>
<td>n+</td>
<td>p-</td>
</tr>
<tr>
<td>3</td>
<td>m+</td>
<td>f+</td>
<td>n-</td>
<td>p-</td>
</tr>
<tr>
<td>4</td>
<td>m+</td>
<td>f-</td>
<td>n+</td>
<td>p-</td>
</tr>
<tr>
<td>5</td>
<td>m+</td>
<td>f-</td>
<td>n-</td>
<td>p-</td>
</tr>
<tr>
<td>6</td>
<td>m+</td>
<td>f+</td>
<td>n+</td>
<td>p+</td>
</tr>
<tr>
<td>7</td>
<td>m+</td>
<td>f-</td>
<td>n+</td>
<td>p+</td>
</tr>
<tr>
<td>8</td>
<td>m+</td>
<td>f-</td>
<td>n-</td>
<td>p+</td>
</tr>
<tr>
<td>9</td>
<td>m-</td>
<td>f-</td>
<td>n-</td>
<td>p-</td>
</tr>
<tr>
<td>10</td>
<td>m-</td>
<td>f-</td>
<td>n-</td>
<td>p+</td>
</tr>
<tr>
<td>11</td>
<td>m-</td>
<td>f-</td>
<td>n+</td>
<td>p-</td>
</tr>
<tr>
<td>12</td>
<td>m-</td>
<td>f-</td>
<td>n+</td>
<td>p+</td>
</tr>
<tr>
<td>13</td>
<td>m-</td>
<td>f+</td>
<td>n-</td>
<td>p-</td>
</tr>
<tr>
<td>14</td>
<td>m-</td>
<td>f-</td>
<td>n+</td>
<td>p-</td>
</tr>
<tr>
<td>15</td>
<td>m-</td>
<td>f+</td>
<td>n+</td>
<td>p+</td>
</tr>
<tr>
<td>16</td>
<td>m-</td>
<td>f+</td>
<td>n-</td>
<td>p+</td>
</tr>
</tbody>
</table>

In the above table, m+ represents a change in stress from the LF-N for masculine forms of adjectives, f+ for feminine, n+ for neuter and p+ for plural. Conversely, m- represents no change in stress for masculine forms, f- for feminine, n- for neuter and p- for plural. Table 5.8, below, gives full examples of this with adjectives which commonly take the short form. Although, mathematically, there are 16 possible combinations, not all of these were found to exist with the adjectives in the sample. Results taken from the Barentsen Corpus, are given below for individual tokens:

218
As is clear from the table above, the majority of the adjectives maintaining their SF in the second half of the 20th century - 75% - have the same stress in both LF-N and SF. This strongly suggests that maintenance of a familiar stress pattern may be key to a lexical item keeping its SF.

However, there is one further group of adjectives which cannot be included in the above table. This group does not have a single fixed stress pattern, in that one or more of the genders may have two possible stress patterns. 49 of these adjectives had two possible stress positions for the plural of the SF, 9 for the neuter, and 2 each for the masculine and feminine. Therefore, a stress alternation for the plural SF is, according to this sample, more than five times more likely than for the neuter, and nearly 25 times more likely than for both the masculine and feminine SF.
This may be due to the relatively low usage of plural SFs. As shown in the figure given from the Corpus study presented in this thesis, plural SFs occur relatively rarely, suggesting that the low frequency of usage for such forms has led to a divergence in possible stress.

402 individual adjectives were sampled from the final quarter of the Corpus. Of these, 146 occurred in the SF in the second half of the 20th century. The rest occurred only as LF-N or LF-I. This is not to suggest that the SF of the other adjectives could or would not be found either within the Barentsen Corpus or any other, but it does suggest that there is still some variation in lexical usage. We could perhaps infer from this that few – if any – SFs are actually dropping out of “possible” usage; that is to say, it is still possible to use the SF of many adjectives, even though these forms may have dropped out of fashion. Similarly, those found may simply have “re-emerged” as SFs of choice.

We can see from the above table that the majority of adjectives found predominantly in the SF in the predicate position have a fairly stable stress pattern – only bol’nøj, dalekij, dolžnyj, nužnyj, pravyj and sklonnyj have a change in stress, along with sčastlivyj, which is alone in changing the stress uniformly throughout the genders and number. Only in bol’nøj and sčastlivyj does the stress change in the masculine, while all the others have a change at least in the feminine. Even when the stress does change, it is a move from first to final vowel, (with the exception of sčastlivyj, which goes from second to first), making the change in stress fairly simple. The greatest amount of stress change occurs in the adjectives found most often in the Corpus – dalekij, dolžnyj and nužnyj, suggesting that frequency of usage has allowed the embedding of these forms into the language without much difficulty. The hypothesis, therefore, would be that an adjective occurring less often than dalekij, dolžnyj and nužnyj and which does not change its stress pattern between LF-N and SF is more likely to maintain its SF. In order to test this hypothesis, we now look at the other adjectives occurring at the end (1951-2000) of the Corpus. Therefore, from a possible 68 changes of stress, we find that there are only 14 changes.
In summary, difficulty in knowing where the stress occurs in the SF may account for its decline. With adjectives such as those listed above, many of which occur frequently throughout the Corpus, we may surmise that frequency of usage has led to familiarity with stress pattern, therefore allowing no confusion as to where the stress may lie.
5.5 DISCUSSION

In this study the total number of SF, LF-N and LF-I adjectives have been analysed, along with a representative sample of the data. When sampling a population it is often impossible or impractical to sample every individual data point. Equally, a sample that is too small may be highly irrelevant and, at worst, lead to theories and explanations that have little or no sound basis. More frequently, a representative sample of the overall dataset is analysed. It is once more worth emphasising the analogy with sampling techniques for opinion polls, which is probably the most widely known example of sampling a population, and is discussed below.

The SF is declining overall, however we can see no clear patterns emerging regarding adjective class. As would be expected, the larger classes (3 and 5) contain the largest numbers of examples, and so we cannot be certain that the propensity for SF in these classes is purely related to the fact that they have a greater number of possible lexical items, rather than being a specific area for SFs. With regard to the group of adjectives favouring the SF, there is no clear decline in usage, suggesting that a process of fossilization of the form may be occurring.

In examining the above graphs, for specific changes to the occurrence of particular verbs, several clear trends are apparent. Where each individual adjective is analysed, trends in SF, LF-N and LF-I changes are variable. This appears to be dependant upon the sample size or, in the case of this study, the number of each adjective returned after searching. For example, inspection of Figures 5.18, 5.19, 5.30 and 5.32 shows, in most cases, clear trends which map onto the overall summation of results. Such results also indicate with greater confidence that the absence of any forms from a particular category would strongly suggest that it is absent as it is not present in the Corpus, and not because it was the results of a limited search or a limited pool of examples. In the case of these verbs, they all have in excess of 100 examples — 509, 124, 217 and 294 examples, respectively. In several cases, such as Figures 5.24, 5.25, 5.26 and 5.28, where there are 77, 71, 98 and 92 examples, respectively, the trends are not as pronounced but are beginning to show the trends expected from previous examinations of the Corpus. In all other cases, where the
numbers of examples are even lower, no clear trends are discernable. These larger subsets may of course “skew” the overall results, but it is more likely that they are more representative of the overall change in verb usage, as a percentage of the dataset, than some of the verbs where very few examples were found. Therefore, in the determination of trends or change in usage, by quantitative or semi-quantitative methods, the sample size available is essential in helping the researcher to obtain a representative sample of the dataset, and in drawing valid conclusions about the data.

It is also clear from an inspection of the Corpus that not all the examples have been analysed in this study. There are clearly cases where very few examples of particular adjectives have been used in this study. This is due to their occurrence in the dataset chosen for investigation in this study. Other examples exist. However, it is not necessary to investigate the usage of these verbs in their entirety. Once a representative sample has been analysed and validated, the same trends – within reasonable bounds of accuracy which will be determined by the chosen methodology – should be apparent. This has been validated by removing what may be considered representative and unrepresentative samples from the overall dataset, and then re-analysing the data.

Following on from the discussion in Section 2.6, the sample size is central in understanding and quantifying the output from such an analysis. Generally, if the sample size is small then the standard deviation of the population can be substituted with the standard deviation of the sample, and the same equation used. This usually means less than 30, although the representative data subsets in this study appear to be significantly larger (greater than 100, in Figures 5.19, 5.20, 5.23 and 5.28 to 5.32). Equation 1 (p. 58) will allow the determination of sample size when the standard deviation of the population or sample is known. If the sample size is less than or equal to 30, the population must be normally distributed and the population standard deviation must be known in order to use Equation 1, which is presented and discussed in Section 2.5.

An underlying assumption to the theory of population sampling is that the data analysed is a genuinely representative and random sample. For example, in election polling a reasonable cross-section of the population is usually picked. If, for example, pollsters
confined their sampled population to a particular subset of the overall population (e.g. by
gender, age, salary, etc.) then an unrepresentative sample would be produced and this
would yield flawed and possibly skewed results. Therefore, in selecting examples from the
Corpus to be analysed, it is appropriate to sample randomly and not base sampling on
previous studies, particularly if they are qualitative in nature. Analysis of Tables 5.2, 5.3
and 5.4 would suggest that the data in this study is randomly distributed, although large
parts of the available data reside with a small number of specific verbs. This should be
considered in the context of the caveats discussed in the previous section.

By taking one verb as an example it is possible to see if Equation 1 (p. 58) is relevant, and
if a particular sample size can be used instead of the whole population of verb occurrences
from an analysis of the Corpus. Using the data presented in Figure 5.18, which has a total
of 509 occurrences, it is estimated that a representative sample would consist of 219 data
points, given a confidence interval of 5% and a confidence level of 95%. Analysing this
subset results in exactly the same trends for SF, LF-N and LF-I change over the time
periods examined, within the confidence limits stated.

Nevertheless, the current analysis of 20 verbs has indicated that qualitative or, at best,
semi-quantitative estimates of change in usage can be estimated with reasonable
accuracy. The most significant parameter would appear to be the sample size, where
approximately 100 data points are required in order to yield representative results. This,
however, might be better estimated through a quantitative sampling method in future
studies. Further, the changes observed in SF, LF-N and LF-I usage over the time period of
the study indicate that a step-change in usage occurs between the second and third time
periods (1851 – 1900, and 1901 – 1950). Regression analysis suggests that these data do
not fit common models and that a more complex statistical approach might be required in
order to discern future changes within this dataset. These qualitative comments should be
fully contextualised by the next chapters, which statistically analyse the data returned from
searches of the Corpus.

The nature of this study has involved the statistical analysis of discrete forms - SF, LF-I and
LF-N – and has used in its description terms of variance, such as the confidence interval.
This, while working with discrete words or forms, indicates that there is a certain amount of variance in how the change occurs, or a certain degree of error implied within confidence limits of that change (i.e. a 95% confidence interval suggests a 95% chance of the answer being within a given range). This study does not overtly examine the mechanism of change; that is, how the form changes, but rather it examines the trends in that change. Implicit in this last point is the possibility of predicting future change — a simple way to do this would be to “predict” the occurrence of form from a period for which the data is known. For example, the “leave-one-out” approach to validation would indicate that the occurrence of form in the first time period (i.e. 1801 – 1850) could be predicted by extrapolation from later data (i.e. 1951 – 2000). However, the limits of validity would be an issue, as would the lack of a single trend observed in the change from SF to LF-I or LF-N. In other words, a simple interpretation or prediction of trends may be misleading or may only be applicable to each particular word, meaning that inferences from one word cannot be extrapolated onto another, as the manner of change (see Chapters 3 and 4) itself cannot be predicted.

Nevertheless, one underlying feature of this work is that variance is stated clearly as part of the results. There are no discrete results, rather a mean and a standard deviation or level of confidence suggesting a “range” in which a result can sit, or the probability that change is statistically significant is carried out within bounds of error. While such variance is relatively simplistic compared to that which underlies the probabilistic approaches to language change, it originates from the same principles and therefore shares common concepts.

However, by randomly taking adjectives from this list and examining how their occurrence changes over time, no clear trend is apparent with respect to frequency. In the case of the two infrequent examples chosen (болен, bole, ‘sick’ and склонен, sklonen, ‘inclined to’) the trends are rather similar, showing an increase in SF in the middle periods of the study, which declines thereafter. However, by taking these examples in isolation it is difficult to determine whether or not this is a significant trend or an artifact associated with the small sample sizes. The examples exhibiting large sample sizes (довolen, dovolen, ‘satisfied’ and должен, dolžen ‘necessary’) show no clear trend. Therefore, by taking these four examples it is clear that frequency may not play a role in the change of form examined in
this study. Rather, as discussed below, other issues, such as the likelihood of change being a gradient function or, simply, the nature of the dataset and examples chosen, may impact on such generalised comments. Whether such trends are artifacts of the sample size may be an important. The size of the sample examined, as discussed earlier in this chapter in the context of whether or not sufficient data have been returned from searching the Corpus to allow a viable comparison of data to be made, is also an important issue in ensuring that such samples are representative and can be reasonably compared.

This may be further contextualised by, for example, Corbett et al. (2001), who investigated frequency and the emergence of linguistic structure. They examined the relationship between frequency and irregularity, suggesting that, while clear links appeared to exist, there was little statistical evidence to support such claims. They commented that the relationship was complex, and was related to the type of frequency and the degree of irregularity observed. In terms of frequency, absolute refers to the lexeme itself, and should include all different forms of the word, whereas relative refers only to the parts that are changed. In the current study, that would suggest that absolute frequency relates to the ratio of LF-I, LF-N or SF with the total of LF-I, LF-N and SF. Rather than the degree of irregularity (except in the case of degrees of regularity, when the stress changes between LF and SF, for example), as used by Corbett et al. (2001) the degree of variance is more appropriate for adjectives. For example, with dolžen and nužen, these occur predominately in the SF in the predicate position. And with 890 examples returned from a search of one time period (1851 – 1900) such frequency of its usage – relative to other time periods and to the other words that the search comprised of – would expect to lead to a more standardised form in the predicate position. This reflects the comment made above, that knowledge of variation must involve knowledge of frequency, which in turn may impact on the morphological changes observed in this study.

It should also be noted that frequency is only one part of language change that supports a probabilistic approach, a subject discussed in Chapter 2. Another factor, for example, is gradience. When considering a gradient-based approach members of categories at the boundaries may exhibit behaviour that ensures an ambiguous classification, at best. This is a phenomena not only found in linguistics, but which is commonly associated with a myriad
of fields where such classification and analysis are considered. Again, in a very general sense, the use of statistical analysis allows us to set an arbitrary boundary, usually demarcated by the confidence interval, $p$, which allows classification of trends as being statistically different or not different. The methodology of employing strata (very much in opposition to the methods of Benson (1953) and Gustavsson (1971), for example, although the latter study was synchronic in design, not diachronic), to classify change is well justified statistically, but may not be so clear in sociological or related fields. In a very simple example, this is illustrated in Figure 4.43. In this sense, the comment that phoneme membership is gradual and not discrete relates to their continuous distribution across a particular phonetic space (Pierrehumbert, 2003). It should be noted that more peripheral members of a "group" may overlap with another group, resulting in shared membership or boundaries where there are not clear examples of discrete classification. Manning (2003) elaborates on this point by suggesting that it may allow modelling of syntactic category membership, a phenomena in which gradience is also apparent. He cites "marginal prepositions" (i.e. concerning, considering, following) as examples that range from being fully verbal to fully prepositional. This would suggest a distribution of space for words within a group – a group that usually overlaps with adjacent groups – where words on the boundaries merge with one another but those at the core remain discrete from neighbouring groups. Thus, categories in linguistics are central to the subject, but categorisation does not assume that categories need to be categorical. A probabilistic approach replaces the discrete quanta of categories with distributions, in which membership is overlapping and gradient. Clearly, therefore, change between any groupings, whether linguistic categories or those groupings assigned for the estimation of morphosyntactic change (as in this project), is gradual and not discrete. In essence, while this might statistically suggest that there are fewer or greater occurrences of the SF in one group (i.e. that representing the time period 1801 to 1850) caution should be exercised in interpreting such data, particularly at the boundary of different groups.

It was also commented that corpus searches do not usually take into account the grammatical nature of the examples returned but instead result in a cline of well-formedness, where some grammatical forms are more common than others (Manning, 2003; Chater and Manning, 2006). However, in the present study such matters were
considered through the process of manual disambiguation of examples returned from searches of the Corpus. Further, these examples were then analysed according to various grammatical criteria (e.g. tense, aspect, etc.). While this is essentially an issue of grammatical frequency, Manning commented that the distinction between grammatical and ungrammatical is often found in the middle of such a cline and removes the ungrammatical from consideration due to lack of frequency. However, such a process is gradual and not discrete and the distinction between what is grammatical and what is ungrammatical is often subjective and arbitrary. Again, a probabilistic approach to language assessment should allow a description of the full continuum between what is, and what is not, grammatically acceptable. Manning also suggested that the gradualness observed in corpus searches is related to grammatical judgments, where speakers readily assign degrees of acceptable grammar, an essentially gradient, and not categorical, classification. This echoes Pierrehumbert's findings that phonological well-formedness of novel words is a gradient function that could potentially be predicted as a function of the probability of the words' subparts (Pierrehumbert, 2003).

Morphological productivity and decomposition have been discussed extensively by Baayen (2003) and Pierrehumbert (2003). For example, it is accepted that some affixes are productive in the sense that they can give rise to new words (e.g. interested and disinterested) whereas other, despite being present in the same form do not. Not all affixes, as Baayen concluded, are equally productive, and are clearly a gradient function. Therefore, classifying affixes as “productive” or “unproductive” is not wholly representative of the words and their productivity, particularly at boundaries. At morpheme boundaries such change is itself gradient and will vary from one word, or affix, to another. This is due to the role of decomposition in speech perception, where more complex words that decompose are represented with strong morphological boundaries, and those that seldom decompose are associated with weak boundaries.

It is interesting how such works relate to this study. It may reflect how able a particular form is to change, phenomena which relates both to the old word and the new word into which it may change. Hence, it may be as much a case of the likelihood of form acquisition as much as disappearance of form. On a wider scale, the ability to change may relate to the
relative natures of SF, LF-I and LF-N in being able to change to another form. For example, glad (rad) cannot exist as LF-I or LF-N, whereas wooden (derevjannyj) can be either LF-I or LF-N but cannot exist in the SF. No such words were included in this study, as they would clearly introduce bias. However, an analogy to Baayen’s study of affixes may be made. Consideration of the middle ground between, for example, the extremes of glad (rad) and wooden (derevjannyj) does not necessarily have to reflect an arbitrary ability to accept a particular form; rather, a degree of gradience might exist between these two extremes where words have a different probability of change in form, rather than a simple “yes” or “no” that will confirm or deny that such a change is possible. Therefore, the likelihood that a word may adopt a form, or the rate at which such a change occurs, is not necessarily quantized, and may be different for different words. While there is little direct evidence for such phenomena in this study, there is a suggestion of it in the different rates, and modes, of change apparent in Chapters 3, 4 and 5, where unclear trends are observed across word classes and where different words show distinct trends in change. The variance in how individual words change their form might be explained by the nature of the individual words and their innate predilection to each change form. Such change is, however, multi-factorial and not readily quantified in a study of this nature. Manning (2003) further blurs the argument/adjunct distinctions by defining them as being gradient in nature. He suggests that dividing verbal dependents into freely occurring adjuncts and subcategorised arguments is often difficult, suggesting that subcategories can be modelled as “a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability”.

Pierrehumbert (2003) determined that negative inference based on sample size was also an important part of the acceptability of gradience. As a sample size increases, so too does the evidence in support of, or against, a particular theory. This is demonstrated by a series of results that show how the acquisition of phonemes and phonological generalisations are continually and gradually changed as the sample increases, suggesting the continual updating of probability distributions based on sample size. While Pierrehumbert outlined such conclusions linguistically, they are commonly associated to others in a wide range of fields. However, the implications for the findings in this study are significant, as they would be in any field. The current study does not represent a definitive conclusion to the question
of short-term morphosyntactic change of predicate adjectives in Russian; rather, it suggests that, with the current methodology and corpus, the results are consistent and validated by the linguistic and statistical methods used. However, there are clear exceptions. In some cases – presented variously across this chapter as well as Chapters 3 and 4 – where there are too few examples of a particular lexical item to evaluate statistically, the work and its conclusions are clearly limited in scope. Only qualitative inferences can be reached from such data. However, it is not outside the realms of possibility that the Corpus will be expanded or that an entirely different corpus becomes available at some point in the future. In these cases a re-analysis of those datasets might yield the same or different conclusions. However, purely in terms of “gradience” of analysis and linguistic endeavour, such improvements are likely to be incremental rather than revolutionary. As was commented on earlier in this chapter, the improvement in methods of searching and disambiguation, and the computerisation of the databases generated from searches will speed up this process, resulting in an iterative improvement to the models developed. In a sense, the very nature of the analysis conducted is gradient.

Now, this is not to dismiss classical linguistics. Rather, the understanding of a probabilistic approach to, in this case, morphosyntactic change, may enhance classical understandings of language change. For example, probabilities also operate at the morpheme level. As discussed above, some affixes are more productive than others (i.e. the probability of use varies) and this forms part of the linguistic knowledge of the speaker. The choice among individuals to use particular affixes that are available will demonstrate a strong bias towards the most probable one, which is itself measured by patterns of occurrence in related words (Baayen, 2003). Hence, the choice of affixes is probabilistic. This is echoed in lexical frequency, where the most frequently chosen words are more likely to change, become irregular or become set in a new form. Hence, word representations, including those between word pairs, are also probabilistic. Finally, this may also relate to syntactic structures, where frequently encountered sentences (or fragments of sentences) are more easily processed than infrequently processed ones. Listeners and readers are influenced by the chance of occurrence of a word or a fragment based on previous exposure. This may influence processing time and understanding, and is involved in disambiguation. Hence, sentence structure is probabilistic.
Therefore, at almost all levels language, and language change, exhibits evidence of probabilistic occurrence rather than occurrence in discrete categories. While such categories are clearly defined, their boundaries are not isolated and their distribution is probabilistic and, in certain cases, bound to overlap. This has clear implications for language change and the statistical methods employed to characterize them.
Chapter 6

Conclusions
The main findings of this study occur in two areas. Firstly, the development and validation of quantitative statistical analysis of linguistic data. Study design, suitable stratification of data and adequate sample size are all essential in order to produce valid results. Secondly, the findings of the study itself. There can be little doubt that the SF is declining. In Chapter 3, we saw that semi-copula verbs are increasingly favouring the LF-I, but choices for the copula verb still diverge. It was also interesting to note that, where the LF-I was not possible (the present tense of the verb to be), the SF was not automatically replaced by the LF-N. Israeli's work (2007) suggests that the instrumental form is now taking on the role of marking a temporary state, therefore where this cannot happen, it would seem that the SF maintains its position. In Chapter 4, we saw that noun type has little bearing on the choice of predicate adjective form, although in Chapter 5, where a specialised group of adjectives were seen to favour the SF, many of that particular group were found predominantly with animate nouns. Chapter 5 also showed that adjectives maintaining their SFs were more likely to share the stress position with their LF-N counterparts.

It has been suggested by previous research that the short form has been decreasing in usage in the Russian language. Therefore, this study focused on corpus-based methods to characterise changes in the Russian predicate adjective over a two hundred year period. In doing so, it used more robust methods than the previous studies. Corpus searches, followed by manual disambiguation of examples returned from those searches, along with appropriate statistical methods, provided a clear indication of patterns of change. This indicated that the short form was decreasing in certain areas, dependent upon the ability of a lexical item to adopt the long form instrumental. This study also demonstrated that the methodology was central to producing valid and linguistically meaningful results. It demonstrated that substantial and statistically significant linguistic (in this case, morphosyntactic) change can occur over a relatively short time period.

Corpus searches are time-consuming and associated manipulations are largely manual in nature. This presents the possibility of errors in the data. It is also a rate-limiting step, along with manual disambiguation, to the analysis of the Corpus. It is recommended that work be undertaken to streamline the operation of the software. This will save substantial amounts
of time and result in more efficient and speedy searches of the Corpus, allowing for more
detailed analysis – qualitative and quantitative – to be undertaken.

Qualitative analysis, as shown in Chapter 2, is extremely useful and should not be
dismissed; nor should statistical methods used subsequently be considered a panacea.
Rather, a combination of both techniques can provide a comprehensive understanding not
only of language change, but of the validity of the results obtained. This is particularly
important with regard to sample size.

Chapter 3 examined verbs using statistical methods. Significant trends were observed, and
the range of available statistical tests was discussed. In general, significant changes in
frequency of occurrence of forms were observed. In most cases (i.e., where there exist
sufficient examples to allow a statistical test to be performed) SF decreases and is
replaced by LF-I. This is echoed in Chapter 4, where nouns were examined, and similar
trends are observed. Where LF-N is the only option for change from SF, little change is
seen, suggesting that changes in LF-N and LF-I are different, and the nature of LF-N
makes it less conducive to change than LF-I. It is not clear why this is the case, but may be
due to restriction of type. Trends are different (bimodal or linear) and seem to depend on
the types examined.

SF is essentially a special form of nominative which has adapted to the predicate position.
There may be a greater overlap or similarity between LF-N and SF than the three
demarcations used in this study tell us, and even that comparisons should be made
between SF/LF-N versus LF-I, rather than SF versus LF-N/LF-I. So, in terms of future work,
one might wish to re-analyse the statistics in this study in such a different context. It should
be remembered, however, that some adjectives are used largely in the SF, some have no
SF but no adjective has SF and LF-N but no LF-I.

With semi copula verbs the instrumental is becoming the form of choice – exact rate and
frequency of occurrence depends on specific examples, shown in each chapter.
The results of this work clearly demonstrate that the relative occurrence of form changes throughout the time period examined. Given a wider background and context — that language changes — these results are not a surprise. However, the current study has identified and quantified particular aspects of language change, and therefore seeks to understand the specific change in language usage, not the mechanism of change. For example, if the frequency of one form changes, or predominates, over another this is not a discrete change but a probabilistic and gradual change. The increase in frequency of words, or forms of words, may result in a distribution of use that differs from its previous incarnation. This is clearly demonstrated by the gradual change in occurrence of form shown in Chapters 3, 4 and 5. This may mirror the probabilistic incremental updating of grammars and lexicons associated with language change. When the frequency of a word or form increases within a population this will happen across a range of individuals, this will be best described by a probabilistic modelling of language change or use. Such a mechanism allows the modelling of different generations (perhaps represented in the current study by different time periods) in a comparative sense, where, for example, one generation uses new variants or forms 10% of the time, and a different (later or earlier) generation uses the variation 50% of the time. Such a model is reasonable in the context of this study, as SF, LF-I and LF-N are always comparatively represented as percentages, suggesting migration of use in a gradual manner rather than an absolute, or quantised, manner.

Use of synchronic variation allows the identification of language change, whether within one language or in a comparative sense with a related language. Once a change has been identified, by whatever means (i.e. comparison of vocabulary or statistical analysis), it can be mathematically analysed to ensure that changes are not random and due to chance. Hence, the use of the somewhat arbitrary p-value of 0.05 in this study (as in many others) to demonstrate compliance or otherwise with the null hypothesis. Ringe (1992) suggests that, even though the possibility of a particular change happening by chance is unlikely, it may actually be representative of a larger series of events, one or more of which are relatively more likely to occur. However, the size of the Corpus examined in this study, and the range of adjectives explored, would suggest that such chance occurrences are not seen as artifacts in this study.
In general, the changes observed in form in this study are consistent and are highly unlikely to be the result of chance. For example, Ringe uses the example of words in two different languages that have the same meaning; the probability of similar occurrences is calculated by multiplying their frequencies (as percentages) together. In the current study, such a technique, while not directly applicable in the form described above, would suggest that the probability of all forms changing (in a gradual sense) from SF to LF as a matter of chance is highly unlikely. This is consistent with language change, which takes place gradually and where innovations (new words, new meanings, new forms) are used at different rates in differing social contexts, suggesting that change overall is gradual, and may occur over various time periods, from years to centuries. An examination of the historical record — as, in this study, by examination of the Corpus — results in changes in observed probabilities being quantified. Further examination of specific changes (in this study, the change from SF to LF) may provide evidence for the manner in which language change is initiated and propagated.

The constant rate hypothesis has been expounded by Kroch (1989), who commented that, as changes occur at the level of abstract grammatical parameters, they change at the same rate in every context. This is independent of whether or not all forms have, at the start of the sampling period, similar distributions; rather, depending on the nature of the starting points the changes may all occur at the same rate, but from different “starting points” in all cases. This, again, is another possible issue with the selection of strata in the data set.

Kroch and co-workers have tested their hypothesis by examining sigmoidal changes in language. Sigmoidal, or S-shaped, trends in language change are well established and understood, and have been discussed previously (cf. Chapter 1). Kroch used the function:

\[ P = \frac{1}{1 + e^{-kx}} \]
where \( P \) is the occurrence of a particular form and \( t \) is time; a logarithmic conversion removes the exponential function and results in a linear function:

\[
\ln \frac{P}{1-P} = k + st
\]

where \( s \) is the gradient, or slope of the line, and \( k \) is the intercept. This function relates the change in form over time and allows the changes observed to be quantified; it also allows – within reasonable bounds – extrapolation beyond the range of the current dataset. This is a technique used in many fields, most notably physics, and allows the potential for reasonable predictions of language change to be made. This point should be considered in the context of the simplified example of extrapolation presented previously, in Chapter 4.

The slope of the line also allows the rate of change to be estimated, in units of percentage occurrence per unit time. Such a regression is rather simplistic, and does not take into account the complex nature of the Corpus and the examples therein. Thus, different weightings and parameters can be added to make such equations more realistic. However, qualitative examination of the data presented in Chapters 3, 4 and 5 would suggest that such an approach is not suited to all the examples examined in this study. The range and complexity of the changes observed (which, while they can be simplified as “the short form decreases in occurrence”) may require more complex modelling in order to consider the different trends observed for different examples. Hence, Figures 4.1 to 4.4 show the changes in frequency of distribution of SF, LF-N and LF-I across the four time periods for a range of examples. In all cases, simple sigmoidal changes are observed, but not consistently. Changes observed, not just in the four Figures mentioned but more broadly in this study, appear to be linear, bimodal or even exponential in nature. This may be an artifact of the data used, and the number of examples – which is inconsistent due to the nature of the Corpus searches – but in the majority of cases the datasets used are larger than those employed by Kroch. This does not suggest that Kroch’s analysis is incorrect – indeed, it is clear that Kroch and colleagues have validated their studies across a range of languages including Old English, British English, Yiddish and Portuguese (references, page 149) but that the model of S-shaped language change may be language, or form, specific. It does, however, suggest that statistical analysis can be used effectively to
characterise the nature of language change and even to predict future language change. This may indeed be more useful than trying to determine the mechanism of change from such data sets.

Tabor (1994) has developed a network model of morphosyntactic change that connects words. Words will connect to similar words and will also be loosely grouped into gradient classical categories, such as verb, noun, or adjective. The relative strengths of syntactic behaviour are therefore dependent on associated words and their proximity in groups or clusters. Clearly, this is related to the nature of the dataset employed and is prone to change, or shift, if the dataset (in this case, the Corpus) is changed or modified. This will result in a gradual adjustment of the model to compensate for the changing nature of the data entered. This is an important consideration as the Corpus is added to and is significant in how the results of this study can be interpreted – highlighting the Corpus as the source of the analysis and its conclusions might seem obvious, but less obvious is that different conclusions – substantially different or only slightly different – may be observed if a different dataset is interrogated. This also has implications for Tabor’s model, as it may, for example, ensure that the connections and clusters formed in one model, based on a particular dataset, will differ if the dataset is changed, changing the relationships within the connectionist network, suggesting frequency-related linkage effects, where a change in one aspect of the language will result in changes in other aspects being made. This is particularly relevant to the current study, and consideration of the changes observed in form.

Irrespective of whether or not language change is sigmoidal over time, such non-linear changes of whatever trend (observed in this study or elsewhere) are important to consider in a wider context. For example, if language change over a particular period is sigmoidal, then it suggests that the change, or changes, begin slowly, accelerate and then decrease towards the end of a particular period. Can one also ensure that the beginning of the sampling period, or of particular strata within the sampling period, occur at the start, middle or end of a particular response curve, S-shaped or otherwise? This point has been considered in Chapter 4, and indeed is relevant to the findings of Chapters 3, 4 and 5. Periodicity in sampling is clearly an issue in such matters – how can one be sure that the
sometimes arbitrary starting point is the "correct" place to start? Do all the studies undertaken by Kroch see an even S-shaped change in language? If the change in language occurs at different times in different studies, or at different rates, then the trend observed may not always be sigmoidal.

Bloomfield (1933) considered that language change will vary from speaker to speaker and will depend on the number of interactions between speakers of the old and new variants. In the early stages of a new language variant coming into existence the rate of change will be small, due to the small number of speakers using the new variant. As time passes, the number of speakers of the new variant will slowly increase, increasing interactions between the old and new variants, at which point the frequency of change will increase. The rate of language change will increase thereafter, reaching a "steady state" when the balance between old and new variants is relatively constant. After this point the rate of change will decelerate, as the frequency of old variants interacting with new variants will decrease as fewer old variants will exist. In time, the frequency should tail off until such a point that stasis, or equilibrium occurs, or the old variant disappears altogether.

This is a classical and somewhat theoretical description of the process of language change. While it expounds a sigmoidal pattern in the rate of change it is relatively simplistic and fails to take into account several key issues. For example, the assumption about the "starting point", discussed above, appears to rely on a point in time where the old variant is present 100% of the time. This fails to consider gradual increments in variation and ties the S-shaped pattern of change into an arbitrary function. As the results of this study have shown, language does change but it does not necessarily change in an exclusively sigmoidal shape within particular sampling intervals. Rather, the consideration of lag times in change, as well as different shapes of change, simply reflects the variance and gradation of change from lexical item to lexical item. Shi (1989) examines language change in Mandarin over a 1,000 year period. Over this period, the majority of the change was contained within a 200 year window. This was attributed to the maintenance of classical styles in the literature, which potentially suppressed or slowed the common acquisition of the new variant. Such apparent punctuated evolution of change could present itself as a sigmoidal change in language. Further reductions in the rate of change towards the end of
the time period examined are attributed by Shi as authors’ attempts to emulate the classical style. Normalisation against the occurrence of a variant in the classical form (for example, 8 occurrences of ye per 1,000 words) suggests a different trend – a sharp rise is observed in usage from the tenth to the twelfth centuries, and no change is observed thereafter. Shi (1989) concludes that conservatism in the written form may mask the gradation, or sharpness, of certain changes in usage by ensuring that change is gradual. This work should be compared to the current study, where the sampling period began at 1800, mapping the widespread introduction of the printing press in Russia. This provides a suitable sociological point at which to begin sampling in this study, and removes a number of concerns discussed above from the current study.

However, Manning (2003) suggests that the Stochastic Optimisation Theory proposed by Boersma and Hayes (Boersma, 1998; Boersma and Hayes, 2001) can evaluate such changes in a predominately sigmoidal model. This model assumes that one constraint in the system changes (rises or falls) at a constant rate throughout the grammar. In their theory, surface forms are chosen based on the satisfaction of constraints whose rankings are normally distributed. Change is therefore slow at the edges of the normal distribution and faster at the centre.

Niyogi and Berwick (1995) derived a model that has different members of the population using different grammars. Learners should decide which grammar to adopt, an assumption that predicts no variation within individuals. Learners draw randomly two utterances by members of the population. Selection of either utterance is based on the support in a particular setting. If the utterances conflict – that is, if they are ambiguous – the more recent utterance prevails. The case was simulated by the use of three parameters governing the constituent order, yielding $2^n$, or in this case $2^3$, being eight, possible grammars. They found that the relationship was in some cases logistic (S-shaped) but that variations in steepness, or the rate of change, were observed. However, in other cases change occurred immediately, while in others it took longer to begin changing. The end was observed as being either the drop-off from a curve or a conventional asymptotic decay. However, even the definition of an “end point” may be subjective and depends on the nature of any stratification in the sample.
This model was then applied to changes observed between Old French (verb-second, V2) to Modern French (VSO) after the criteria established by Clark and Roberts (1993), where five binary parameters are employed, resulting in \(2^5\), or 32, possible grammars. Change and loss of V2 depends on the criteria employed at the start of the test. Specifically, if small numbers of VSO speakers are used from the start of the test language change is faster than if none were initially employed.

The model proposed by Niyogi and Berwick is deterministic in nature and is based on an infinite population of agents and generations that do not overlap. Briscoe (2000) extended the model to a scenario where the population was small and finite. In doing so, substantially different results were found between both models. If two competing grammars are initially equally distributed and yield equal proportions of ambiguous sentences, then in the infinite population model of Niyogi and Berwick equilibrium between both grammars will occur, with half the learners adopting one grammar and the other half adopting the other grammar. By contrast, in a finite and smaller population such equilibrium will be difficult to achieve as the probability is low that exactly half the population will adopt one grammar. Such a scenario will see one grammar predominate over the other. In this case it becomes increasingly unlikely that the dominant grammar can maintain its advantage. For example, if a grammar containing a reasonable proportion of unambiguous sentences is used by 100% of the population, this will eventually see some members of the population use the other grammar. Depending on the degree of ambiguity, the language will eventually settle into a pattern where usage oscillates from one grammar to the other.

The above examples try to fit the proposed sigmoidal response to language change. In most cases they do so in specific cases where a small number of words are used, or where clear assumptions are established prior to any modelling. Such approaches sit apart from the findings of the current study, ostensibly due to the mode of analysis and the scope of this study. The scope of this study is larger than those discussed above, particularly when considering the number of words changing. This is interesting as the results in Chapters 3, 4 and 5 clearly demonstrate that some forms are changing in a sigmoidal manner, but that others are changing in linear or bimodal manners. While exact classifications of change are
arbitrary and relate to the sampling interval of the study as much as anything else, it is clear, particularly from Shi's work, that the changes observed in usage are not necessarily sigmoidal.

It is important therefore to analyse in detail the nature of the dataset, the type of analysis, the nature of any stratification and the range of examples being studied. All may have an effect on the nature of change. For example, a wide range of examples are considered in this study. Some show S-shaped change, some do not. If language change is S-shaped, then its absence in these cases may relate to the size of the sample, the size of the time period examined and, consequently, may reflect on the rate of change; specifically, that the rate of language change differs for each example used and that this is reflected in observing merely a snapshot of the overall picture. While Shi examined a one thousand year period and saw rapid change over a two hundred year period but little else, it is not unrealistic to suggest that the opposite may be the case for some words examined in this study. In essence, such an argument implicitly returns to the fundamental subjectivity of the analysis of each study, and the analysis of particular words. It may suggest that the establishment of an arbitrary sampling period suits some lexical items more than others, i.e. the special group of adjectives analysed in Chapter 5 (those most frequently occurring in the SF) and those words described by Pagel et al. (2007) as being more or less susceptible to change based on frequency of usage), in terms of discerning linguistically plausible trends. While Manning commented that a "realistic and complete model of how changes spread remains to be implemented" it should be considered that such a goal – a holistic model for all language change – simply may not be possible as it may not reflect the nature of the grammars being modelled.

Interestingly, such change has also been considered in the context of available grammars. For example, Richards (1997) examined changes in the Australian language Lardil (as Old Lardil and New Lardil). He investigated changes that were not just lexical, but grammatical. Lardil is being replaced by English in everyday use. However, Richards comments that the changes are not due to the presence of English but due to the absence of Old Lardil to compete with New Lardil. This under-transmission of morphosyntax is dependent on the syntactic sensitivities of the changes that would not be expected if Lardil was simply
adopting English morphosyntax. Therefore, the presence or absence of an old form of the grammar may have as significant a role in its evolution as the presence of an entirely new language.

The modelling of language change, in this study and in the work of others described above, may lead to a mechanistic understanding of how language changes. They shed light on learning mechanisms as well as the social function of the language. As the application of probabilistic methods in language change are applied the models developed can provide validation, within the framework of the model and the data used, of particular linguistic theories, such as the S-shaped pattern associated with language change. Adding to such models by addition of probabilistic methodologies enhances their realism and applicability, and allows the selection of models — including particular aspects of probabilistic behaviour — based on their accuracy to existing qualitative information. Hence, a model or a methodology that fits the “facts” very well is preferred to those that do not. From such approaches reasonably predictive models will be developed, potentially allowing the prediction of language change.

The consideration of probability in structuralist and generative theories of morphology is something that has only recently grown in interest. This is in contrast to statistical research in language variation which has been widely explored. Baayen (2003) reviews the background to this change in detail, and in particular considers how linguistics research has been changed by improvements in computing power. He focuses on connectionist past tense artificial neural network (ANN) models of McClelland and Rumelhart (1987, cited in Baayen, 2003) and their ability to predict the present tense.

Such models were widely criticised, possibly due to the concern regarding the prediction of future language change by the imposition of an often rigid and potentially limited mathematical theory. For example, Pinker and Prince (1988) suggested that the McClelland and Rumelhart ANN was massively flawed in its predictions of regular and irregular forms. They suggest that the ANN model has no explicit form, that the model lacks any specific rules and contains only a set of neuron-style units which represent trigrams of phonetic features of the stem, a set of units which stands for triagrams of phonetic features
of the past form and an array of connections between the two sets of units whose strengths are modified during learning. Pinker and Prince suggest that McClelland and Rumelhart's assertion that linguistic rules are nothing more than "approximate fictions" and that the real causal processes in language use and acquisition must be characterised as the transfer of activation levels among units and the weights of their connections. Pinker and Price analysed McClelland and Rumelhart's model and found that, in their analysis, it could not represent certain words or learn many rules (or, indeed, learn rules not found in any human language). Further, it cannot explain morphological and phonological regularities, nor can it explain the differences between regular and irregular forms. It incorrectly explains development phenomena proposed by McClelland and Rumelhart and, finally, it fails to predict the tenses at all. Pinker and Prince's article is one of many in this field (i.e. MacWhinney and Leinbach, 1991; Seidenberg and Hoeffner, 1998; Juola and Plunkett, 2000) and it serves to highlight a key issue in modelling – the limitations of the model may very much depend on the input, which will clearly influence the output. Pinker and Prince's analysis of the ANN model clearly demonstrates this by suggesting that the model is limited in scope. The interpretation of the output and accuracy of the model is another issue, but any model is clearly only as good as the data input and the method used for analysis. Evolution of models may be reflected in a change in Pinker's stance; for example, in early responses to the connectionist ANN studies (Pinker and Prince, 1988; Pinker, 1991) the connectionist approach was completely dismissed. Later (Pinker, 1997, 1999) his position is modified by suggesting the possibility of storage of irregular verbs in associative memory while maintaining his "words and rules" claim. Marcus (2001) is more accepting of connectionist models, but highlights their clear limitations as they cannot explain particular datasets associated with revealing the apparently symbolic nature of human language processing.

The above studies suggest one possible future for linguistics. The current study, while adopting some of the above methodology has as much as anything else utilised the concepts of gradience and probabilistics in language change. Sound and validated methodology clearly underpin any corpus study of value, and such investigations should encompass techniques from fields such as statistics in order to enhance and develop the knowledge gained from corpus studies. In the case of this study, specific elements of
morphosyntactic change have been identified, as has the ability to observe dramatic language change over a relatively short period of time.

This study is linked with several similar studies cited throughout this document, particularly those by Benson (1953) and Gustavsson (1976). While the current study has made methodological improvements over the previous works it is part of the continuum of corpus studies. Therefore, this work may in the future be viewed in a similar context. Rapid changes in computational resources and the application of statistics to this field will enhance the ability of researchers to achieve faster results with larger numbers of examples. Therefore, while technology has, and will continue, to shape how corpus linguistics develops as a field the focus will be placed more into linguistic understanding than method development. For example, this would allow researchers the ability to focus in more detail on changes within, and across, specific genres.

Further, the ability of certain words to adopt particular forms is an area of interest, especially in the context of where such change might be limited – this study has shown that under certain conditions the SF does not decrease if the only option is the LF-N (i.e. the LF-I is not possible, for example with a zero copula). Finally, while the initial predictive studies in this document were reasonably promising but showed the potential pitfalls of such methods, substantial work needs to be conducted in this field in order to fully evaluate the potential for prediction of language change.
List of texts used in this study
(Where an English translation exists, this title has been used. Otherwise, all translations are my own)

1801-1850

Aksakov K.S.
Облако (Oblako) ‘The Cloud’
Вальтер Эйзенберг (Val'ter Ejzenberg) ‘Walter Eizenberg’

Aksakov
Охотная страсть (Oxotnaja strast’) ‘A Passion for Hunting’

Baratynskij
Перстень (Peresten’) ‘The Ring’

Bestuzhev-Marlinskij
Аммалат-Бек (Ammalat-BeK) ‘Ammalat-Bek’
Вечер на бивуаке (Večer na bivuake) ‘Evening at the Camp’
Вечер на кавказских водах в 1824 году (Večer na kavkazskix vodax v 1824 godu) ‘An Evening on Caucasian Waters in 1824’
Второй вечер на бивуаке (Vtoroj večer na bivuake) ‘Second Evening at the Camp’
Замок Эйзен (Zamok Ejzen) ‘Castle Eisen’
Испытание (Ispytanie) ‘The Trial’
Изменник (Izmennik) ‘The Traitor’
Латник (Latnik) ‘The Armour-clad Man’
Лейтенант Белозор (Lejtenant Belozor) ‘Lieutenant Belozor’
Замок Венден (Zamok Venden) ‘Castle Venden’
Замок Неугаузен (Zamok Nejgauzen) ‘Castle Neuhausen’
Ревельский турнир (Revel’skij turnir) ‘The Tournament at Revel’
Роман и Ольга (Roman i Ol'ga) ‘Roman and Olga’
Страшное гаданье (Strašnoe gadan'e) ‘The Terrible Fortune-Telling’

Davydov
Дневник партизанских действий 1812 года (Dnevnik partizanskix dejstvi 1812 goda) ‘The Diary of Guerrilla Warfare, 1812 [Partisan’s Diary]’

Durova
Кавалерист-девица (Kavalerist-devica) ‘The Cavalry Maiden’

Gan
Идеал (Ideal) ‘The Ideal’

Glinka
Очерки Бородинского сражения (Воспоминания о 1812 го)] (Očerki Borodinskogo sraženija (Vospominanija o 1812 go)] ‘Essays on the Battle of Borodino (Memoirs of 1812)’
Письма русского офицера (Pisma russkogo oficera) ‘Letters of a Russian Officer’

Gogol
Мертвые души (Mertvye duši) ‘Dead Souls’
Миргород (Mirgorod) ‘Mirgorod’
Ревизор (Revizor) ‘The Government Inspector’
Женитьба (Ženit'ba) ‘The Marriage’
Невский проспект (Nevskij prospekt) ‘Nevskij Prospekt’
Нос (Nos) ‘The Nose’
Портрет (Portret) ‘The Portrait’
Шинель (Šinel’) ‘The Overcoat’
Коляска (Koljaska) ‘The Pram’
Записки сумасшедшего (Zapiski sumasšedšego) ‘Diary of a Madman’
Рим (отрывок) (Rim (otryvok)) ‘Rome (fragment)’
Вечера на хуторе близ Дikan’ки (Večera na xitore bliz Dikan’ki) ‘Evenings on a Farm Near Dikan’ka

Karamzin
Бедная Лиза (Bednaja Liza) ‘Poor Liza’
История государства Российского. Том V (Istorija gosudarstva Rossijskogo. Tom V) ‘History of the Russian State, Volume 5’
История государства Российского. Том VIII, (Istorija gosudarstva Rossiskogo. Tom VIII) ‘History of the Russian State, Volume 8’

Киреевский
Опал (Opal) ‘Opal’

Лащенников
Новобранец 1812 года (Novobranec 1812 goda) ‘Novobranec, 1812’

Лермонтов
Герой нашего времени (Geroj našego vremen’i) ‘Hero of our Time’
Княгиня Лиговская (Knjaznja Ligovskaja) ‘Princess Ligovskaja’
Я хочу рассказать вам (Ja xoču rasskazat’ vam) ‘I want to tell you’
Вадим (Vadim) ‘Vadim’
Штосс (Štoss) ‘Štoss’

Мельгунов
Кто же он? (Kto že on?) ‘Who is it?’

Павлов
Демон (Demon) ‘Demon’
Ятаган (Jatagan) ‘The Scimitar’

Погодин
Дубравка (Dubravka) ‘Dubravka’
Черная немочь (Černaja nemoć’) ‘Black Sickness’

Погорелский
Двойник или мои вечера в Малороссии (Dvojnik ili moi večera v Malorossii) ‘The Double, or My Evenings in Ukraine’
Черная курица, или Подземные жители (Černaja kurica, ili Podzemnye žiteli) ‘The Black Hen, or Living Underground’

Пушкин
Капитанская дочка (Kapitanska dačka) ‘The Captain’s Daughter’
Повести Белкина (Povesti Belkina) ‘The Tales of Belkin’
Арап Петра великого (Arap Petra velikogo) ‘The Moor of Peter the Great’
Гости съездались на дачу: Роман в письмах (Gosti s’ezealis’ na daču: Roman v pis’max) ‘The Guests Gathered at the Dacha: A Novel in Letters’
История села Горокина (Istorija sela Goroxina) ‘The Story of the Village of Goroxino’

249
На углу маленькой площади... (Na ugle maien'koj ploščad...i...) ‘In the Corner of a Small Square’

Рославлев (Roslavlev) ‘Roslavlev’

Дубровский (Dubrovskij) ‘Dubrovskij’

Пиковая дама (Pikovaja dama) ‘The Queen of Spades’

Кирджали (Kirdžali) ‘Kirdzhali’

Египетские ночи (Egipetskija noči) ‘Egyptian Nights’

Повесть из римской жизни (Povest' iz rimskoj žizni) ‘A Tale from Roman Life’

Мария Штопинг (Maria Stoping) ‘Maria Stoping’

Отрывки и наброски (Otryvki i nabroski) ‘Fragments and Sketches’

Планы ненаписанных произведений (Plany nenapisannyh proizvedenij) ‘Plans of Unwritten Works’

Путешествие в Эрзерум во время похода 1829 года (Putešestvie v Erzerum vo vremia poxoda 1829 goda) ‘Journey to Erzerum During the Campaign of 1829’

Saltykov-Ščedrin

Господа Головлевы (Gospoda Golovlevy) ‘The Golovlev Family’

Senkovskij

Превращение гролов в книги и книг в головы (Prevaščenie grolov v knigi i knig v golovy) ‘The Conversion of the Grols to Books and of Books to Lives’

Somov

Приказ с того света (Prikaz s togo sveta) ‘The Order from the Other World’

Киевские ведьмы (Kievskie ved'my) ‘Witches of Kiev’

Кикимора (Kikimora) ‘Kikimora’

Titov

Уединенный домик на Васильевском (Uedinennyj domik na Vasil'evskom) ‘A Secluded House on Vasil'evsky Island’

Zagoskin

Вечер на Хопре (Večer na hopre) ‘An Evening on the Hopra’

Рославлев, или Русские в 1812 году (Roslavlev, ili Russkie v 1812 godu) ‘Roslavlev, or Russians in 1812’
1851-1900

Černyshevskij
“Что делать?” (“Что делать?”) ‘What is to be done?’
Čekhov
Три сестры (Tri sestri) ‘Three Sisters’
Чайка (Čajka) ‘The Seagull’
Дядя Ваня (Djadja Vanja) ‘Uncle Vanja’
Вишневый сад (Višnevij sad) ‘The Cherry Orchard’
Сочинения. Том второй. 1883-1884 (Sočinenija. Tom vtoroj. 1883-1884) ‘Essays, Vol. 2 1883-1884’
Сочинения. Том третий. 1884-1885 (Sočinenija. Tom tretij. 1884-1885) ‘Essays, Vol. 3 1884-1885’
Сочинения. Том четвертый. 1885-1886 (Sočinenija. Tom četvertyj. 1885-1886) ‘Essays, Vol. 4 1885-1886’
Сочинения. Том пятый. 1886 (Sočinenija. Tom pjetyi. 1886) ‘Essays, Vol. 5 1886’
Сочинения. Том шестой. 1887 (Sočinenija. Tom šestoj. 1887) ‘Essays, Vol. 6 1887’

Danilevskij
Сожженная Москва (Sožžennaja Moskva) ‘Burnt Moscow’

Dostoevskij
Бесы (Besy) ‘The Devils’
Братья Карамазовы (Brat'ja Karamazovy) ‘The Brothers Karamazov’
Идиот (Idiot) ‘The Idiot’
Игрок (Igrok) ‘The Gambler’
Подросток (Podrostok) ‘The Raw Youth’
Преступление и наказание (Prěstuplenie i nakazanie) ‘Crime and Punishment’
Бедные люди (Bednye ljudi) ‘Poor Folk’
Белые ночи (Belye noći) ‘White Nights’
Крокодил (Krokodil) ‘The Crocodile’
Дядюшкин сон (Djadjuškin son) ‘The Uncle’s Dream’
Двойник (Dvoinik) ‘The Double’
Хозяйка (Xozjajka) ‘The Landlady’
Маленький герой (Malen’kij geroj) ‘A Little Hero’
Записки из мертвого дома (Zapiski iz mertvogo doma) ‘Notes from the House of the Dead’
Неточка Незвanova (Netočka Nezvanova) ‘Netochka Nezvanova’
Записки из подполья (Zapiski iz podpol’ja) ‘Notes from the Underground’
Скверный анекдот (Skvernyj anekdot) ‘A Nasty Story’
Сон смешного человека (Son smešnogo čeloveka) ‘Dream of a Ridiculous Man’
Вечный муж (Večnyj muž) ‘The Eternal Husband’
Село Степанчики и его обитатели (Selo Stepančikovo i ego obitateli) ‘The Village of Stepanchikovo and its Inhabitants’
Унизенные и оскорблённые (Unižennye i oskorblennyje) ‘The Humiliated and the Insulted’

Garšin
Красный цветок (Krasnij cvetok) ‘The Red Flower’

Gončarov
Обломов (Oblomov) ‘Oblomov’

Kuprin
Синяя звезда (Sinjaja zvezda) ‘Blue Star’
Гамбринос (Gambrinos) ‘Gambrinus’
Колесо времени (Koleso vremeni) ‘The Wheel of Time’
Чудесный доктор (Čudesnyj doktor) ‘A Wonderful Doctor’
Дознание (Doznanie) ‘The Inquiry’
Олеся (Olesja) ‘Olesja’
Пиратка (Piratka) ‘The Little Pirate’
Штабс-капитан Рыбников (Štabs-kapitan Rybnikov) ‘Staff Captain Rybnikov’
На глухарей (Na gluxarej) ‘Hunting the Woodgrouse’

Leskov
Некуда (Hekuda) ‘No Way Out’
На ножах (Na nožax) ‘At Daggers Drawn’
Соборяне (Soborjane) ‘Cathedral Folk’
Очарованный странник (Ocharovannyy strannik) ‘The Enchanted Wanderer’

Писемский
Взбаламученное море (часть четвертая) (Vzbalamučennoe more (čast' četvertaja)) ‘The Unsettled Sea (Part Four)’
Взбаламученное море (часть первая) (Vzbalamučennoe more (čast' pervaja)) ‘The Unsettled Sea (Part Five)’

Соловьёв
История России с древнейших времен (Istorija Rossii s drevnejšix vremen) ‘A History of Russia from the Earliest Times’

Толстой
Альберт (Albert) ‘Albert’
Анна Каренина (Anna Karenina) ‘Anna Karenina’
Война и мир (Vojna i mir) ‘War and Peace’
Воззвание (Vozzvanie) ‘A Proclamation’
Воскресение: часть первая (Voskresenie: čast' pervaja) ‘Resurrection: Part One’
Воскресение: часть вторая (Voskresenie: čast' vtoraja) ‘Resurrection: Part Two’
Воскресение: часть третья (Voskresenie: čast' treťaja) ‘Resurrection: Part Three’
Два гусара (Dva gusara) ‘Two Hussars’
Декабристи (Dekabristy) ‘The Decembrists’
Детство (Detstvo) ‘Childhood’
Из кавказских воспоминаний (Iz kavkazskix vospominanij)
‘Recollections from the Caucasus’
Казаки (Kazaki) ‘The Cossacks’
Люцерн (Ljucern) ‘Lucerne’
Метель (Meteľ) ‘The Snowstorm’
Набег (Nabeg) ‘The Raid’
Отец Сергий (Otec Sergij) ‘Father Sergius’
Отрокчество (Otrčestvo) ‘Boyhood’
Отрывки рассказов из деревенской жизни (Otryvki rassказov iz derevenskoj žizni) ‘Stories from a Country Life’
Поликушка (Polikuška) ‘Polikuša’
После бала (Posle bala) ‘After the Ball’
Путь жизни (Put’ žizni) ‘Way of Life’
Рубка леса (Rubka lesa) ‘The Wood-felling’
Севастополь в августе 1855 года (Sevastopol’ v avguste 1855 goda) ‘Sebastopol in August 1855’
Севастополь в мае (Sevastopol’ v mae) ‘Sebastopol in May’
Семейное счастье (Semejnoe s chastie) ‘Family Happiness’
Смерть Ивана Ильича (Smert’ Ivana Il’icha) ‘The Death of Ivan Ill’ich’
Хаджи-Мурат (Xadžu-Murat) ‘Hadji-Murat’
Юность (Junost’) ‘Youth’

Turgenev

Фауст (Faust) ‘Faust’
Бежин луг (Bežin lug) ‘Bežin Meadow’
Вешние воды (Vesnie vody) ‘The Torrents of Spring’
Дворянское гнездо (Dvorjanskoе gnezdo) ‘A Nest of Gentlefolk’
Дым (Dym) ‘Smoke’
Клара Милич (Klara Milich) ‘The Mysterious Tales (Klara Milich)’
Накануне (Nakanune) ‘On The Eve’
Новь (Nov) ‘Virgin Soil’
Отцы и дети (Otcy i deti) ‘Fathers and Sons’
Первая любовь (Pervaja ljubov’) ‘First Love’
Рудин (Rudin) ‘Rudin’
1901-1950

Ageev
Роман с кокаином (Roman s kokainom) ‘A Novel with Cocaine’

Ajtmatov
После сказки (Posle skazki) ‘After the Fairytale’
Пегий пес, бегущий краем моря (Pegii pes, beguščij more) ‘The Spotted Dog Running at the Edge of the Sea’
Ранние журавли (Rannie žuravli) ‘Early Cranes’

Andreev
Жизнь Василия Фивейского (Žizn’ Vasilija Fiejskogo) ‘The Life of Vasilii Fivejskij’
Красный смех (Krasnij smex) ‘Red Laughter’
Иуда Искариот (Iuda Iskariot) ‘Judas Iscariot’

Babel
Миниатюры (Miniatjury) ‘Miniatures’
Конармия (фрагменты) (Konarmija (fragmenty)) ‘Cavalry (Fragments)’
Переход через Збруч (Perexod čerez Zbruč) ‘Crossing the Zbruc’
Костел в новограде (Kostel v novogradu) ‘The Church at Novograd’
История моей голубятни (Istorija moej golybijatni) ‘The Story of my Dovecote’

Bulgakov
Белая гвардия (Belaja gvardija) ‘The White Guard’
Мастер и Маргарита (Master i Margarita) ‘The Master and Margarita’
Роковые яйца (Rokovye jajca) ‘The Fatal Eggs’
Собачье сердце (Sobač’e serdce) ‘Heart of a Dog’
Театральный роман (Teatral’nyj roman) ‘A Theatrical Novel’
Записки юного врача (Zapiski junogo vrača) ‘Notes of a Country Doctor’

Bunin
Танька (Tan’ka) ‘Tanka’
Поздней ночью (Pozdejn noć’yu) ‘Late at Night’
Антоновские яблоки (Antonovskie jabloki) ‘Antonov Apples’
Туман (Tuman) ‘Fog’
Тишина (Tišina) ‘Silence’
Заря всю ночь (Zarja vsju noč’) ‘Dawn All Night’
У истока дней (U istoka dnej) ‘The Well of Days’
Деревня (Derevnja) ‘The Village’
Сукодол (Suxodol) ‘Suxodol’
Снежный бык (Snežnij byk) ‘The Snow Bull’
Хорошая жизнь (Xorošaja žizn’) ‘The Good Life’
Сверчок (Sverčok) ‘The Cricket’
Море богов (More bogov) ‘Sea of the Gods’
Захар Воробьев (Zaxar Borob'ev) ‘Zaxar Borobév’
Последнее свидание (Poslednee svidanie) ‘Last Meeting’
Личарда (Ličarda) ‘Ličarda’
Лирник Rodion (Lirnik Rodion) ‘Lirnik Rodion’
Чаша жизни (Čaša žizni) ‘Cup of Life’
Братья (Brat'ja) ‘Brothers’
Клаша (Klaša) ‘Klaša’
Грамматика любви (Grammatika ljubvi) ‘Grammar of Love’
Господин из Сан-Франциско (Gospodin iz San-Francisko)
‘The Gentleman from San Francisco’
Казимир Станиславович (Kazimir Stanislavovič) ‘Kazimír Stanislavovič’
Косцы (Koscy) ‘The Mowers’
Преображение (Preobraženie) ‘Transformation’
Лапти (Lapti) ‘Bast Sandals’
Книга (Kniga) ‘The Book’
Митина любовь (Mitina ljubov) ‘Mitya’s Love’
Солнечный удар (Solnechnij udar) ‘Sunstroke’
Ида (Ida) ‘Ida’
Кавказ (Kavkaz) ‘The Caucasus’

256
Дочери Каина (Daughters of Cain)
Путешествие в страну эфира (A Trip to the Ether)

Il'f & Petrov
Одноэтажная Америка (One-storeyed America)
12 стульев (The 12 Chairs)
Золотой теленок (The Golden Calf)

Inber
Смерть Луны (Death of the Moon)

Ivanov, G
Распад атома (Splitting the Atom)

Ivanov, V
Голубые пески (Azure Sands)
Возвращение Будды (The Return of the Buddha)

Kaverin
Два капитана (Two Captains)
Девять десятых судьбы (Nine-tenths of Fate)
Пятый странник (The Fifth Wanderer)

Lagin
Старик Хоттабыч (Old Man Xottabyč)

Leonov
Барсуки (Badgers)
Бегство Мистера Мак-Кинли (The Flight of Mister MacKinlie)

Nabokov
Посещение музея (The Visit to the Museum)
Нежить (The Undead)
Бритва (The Razor)
Драка (The Fight)
Слово (The Word)
Случайность (A Matter of Chance)
Кэмбридж (Cambridge (essay))
Круг (Krug) ‘The Circle’
Король (Korole) ‘Kinglet’
Тяжелый дым (Tjaželyj dym) ‘Torpid Smoke’
Набор (Nabor) ‘Recruiting’
Лик (Lik) ‘The Face’
Истребление тиранов (Istreblenie tiranov) ‘Tyrants Destroyed’
Василий Шишков (Vasili Šiškov) ‘Vasili Shishkov’
Адмиралтейская игла (Admiraltejskaja igla) ‘The Admiralty Spire’
Облако, озеро, башня (Oblako, ozera, bašnya) ‘Cloud, Lake, Tower’
Уста к устам (Usta k ustem) ‘Lips to Lips’
Рождественский рассказ (Poždestvenskij rasskaz) ‘The Christmas Story’
Весна в Фиальте (Vesna v Fial’te) ‘Spring in Fialta’
Юбилей (Esse) (Jubilej (éssé)) ‘Jubilee’

Pasternak
Детство Льюерс (Detstvo Ljuvers) ‘The Childhood of Luvers’
Доктор Живаго (Doktor Zivago) ‘Doctor Zhivago’

Pil’njak
Третья столица (Tret’ja stolica) ‘The Third Capital’
Три брата (Tri brata) ‘Three Brothers’

Platonov
Чевенгур (Čevengur) ‘Chevengur’
Ювенильное море (Juvenil’noe more) ‘The Jewel Sea’
Котлован (Kotlovan) ‘Foundation Pit’
Фро (Fro) ‘Fro’
Счастливая Москва (Sčastlivaja Moskva) ‘Happy Moscow’
Семен (Semën) ‘Simeon’
Сокровенный человек (Sokrovennyj čelovek) ‘The Hidden Man’

Prišvin
Голубиная книга (Golubinaja kniga) ‘The Pigeon Book’
Кошечева цель (Koščeeva cel) ‘The Chain of Koščei’
Šaginjan
Перемена (Peremenia) ‘Change’
Sergeev-Censkij
Балка (Valja) ‘Valja’
Лерик (Lerik) ‘Lerik’
Капитан Коняев (Kapitan Konjaev) ‘Captain Konjaev’
Сад (Sad) ‘The Garden’
Šoloxov
Нахаленок (Naxalenok) ‘The Impertinent Man’
Švarc
Дракон (Drakon) ‘Dragon’
Тень (Ten) ‘The Shadow’
Обыкновенное чудо (Obyknoonnoe čudo) ‘An Ordinary Miracle’
Золушка (Zoluška) ‘Cinderella’
Tèffi
О Дневнике (O Dnevniike) ‘About a Diary’
Tolstoj, A
Аэлита (Aëlita) ‘Aelita’
Гиперболоид инженера Гарина (Giperboloid inženera Garina) ‘Engineer Garin’s Hyperboloid’
Золотой ключик, или приключения Буратино (Zolotoj ključik, ili priključenija Buratino) ‘The Golden Key, or the Adventures of Pinocchio’
Петр Первый (Petr Pervyj) ‘Peter I’
Семь дней, в которые был ограблен мир (Sem’ dnej, v kotorye byl ograblen mir) ‘Seven Days, in which the Earth was Robbed’
Хождение по мукам (книга 1) (Xoždenie po mukam (kniga 1) ‘The Road to Calvary (Book1)’
Хождение по мукам (книга 3) (Xoždenie po mukam (kniga 3) ‘The Road to Calvary (Book 2)’
Tynjanov
Подпоручик Киже (Podporučik Kiže) ‘Second Lieutenant Kiže’
Xarms
Анекдоты, приписываемые Даниилу Хармсу (Anekdoty, pripisyvaemye Daniilu Xarmsu) ‘The Anecdotes of Daniil Xarms’
Рассказы и повести (Rasskazy i povesti) ‘Stories and Tales’
Старуха (Staruxa) ‘The Old Woman’

Зайцев
Голубая звезда (Golubaja zvezda) ‘Blue Star’

Замятин
Мы (My) ‘We’
На куличках (Na kuličkax) ‘At the End of the World’

Зощенко
Избранные рассказы для детей (izbrannye rasskazy dlja detej) ‘Selected Stories for Children’
1951-2000

Aksenov
Апельсины из Марокко (Apel’siny iz Marokko) ‘Oranges from Morocco’
Ожог (Ožog) ‘The Burn’
Пора, мой друг, пора (Pora, moj drug, pora) ‘It is Time, My Friend, It is Time’
Сюрпризы (Surprizy) ‘Surprises’
Катапульта (Katapulta) ‘Catapult’
Перемена образа жизни (Peremená obraza žizni) ‘Change of a Way of Life’
Завтраки сорок третьего года (Zabtrakí sorok tret’ego goda) ‘Breakfast during 1943’
Папа, сложи! (Papa, složi!) ‘Dad, Give Up!’
На полпути к луне (Ha polputi k lune) ‘Halfway to the Moon’
Мой дедушка — памятник (Moj deduška — pamjatnik) ‘My Grandfather is a Monument’
Любителям баскетбола (Ljubiteljám basketbola) ‘For Basketball Amateurs’
Звездный билет (Zvezdný billet) ‘Ticket to the Stars’
Остров Крым (Ostrov Krym) ‘The Island of Crimea’
Круглые сутки non-stop (Kryglye sutki non-stop) ‘Non-stop Around the Clock’

Aleškovskij
Кенгуру (Kenguru) ‘Kangaroo’
Николай Николаевич (Nikolaj Nikolaevič) ‘Nikolaj Nikolaević’
Рука (Ruka) ‘The Hand’

Antonov
Анкета (Anketa) ‘The Application Form’

Aržak
Говорит Москва (Govorit Moskva) ‘Moscow Is Speaking’

Averčenko
Автобиография (Avtobiografija) ‘Autobiography’

Belov
Привычное дело (Privyčnoe delo) ‘An Ordinary Affair’
На Росстанном холме (Na Rosstannom xolme) ‘On Rosstan Hill’
За тремя волоками (Za tremja volokami) ‘Beyond the Three Portages’
Ко (Kon) ‘The Horses’
Тезки (Tezki) ‘The Namesakes’
Мальчики (Mal’čiki) ‘The Boys’
Рассказы о всякой живности (Rasskazy o vsjakoj živnosti) ‘Stories about Poultry of All Kinds’
Моздокский базар (Mozdokskij bazaar) ‘The Mozdokskij Bazaar’
За дальным меридианом (Za dal’nim meridianom) ‘Beyond the Far Meridian’
Плотнице рассказы (Plotniciie rasskazy) ‘A Carpenter’s Stories’
Воспитание по доктору Споку (Vospitanie po doktoru Spoku) ‘Training According to Dr Spock’
Свидание по утрам (Svidanie po utram) ‘Meeting in the Morning’
Дневник нарколога (Dnevnik narkologa) ‘The Diary of a Drug Addict’
Чок-полу́чок (Čok-polučok) ‘Čok-polučok’

Bitov
Жены не́т дома́ (Ženy net doma) ‘The Wife is not at Home’

Dovlatov
Встре́тились, поговори́ли (Vstretilis’) ‘We met, and we talked’
Дальше́ (Dal’še) ‘Further’
Иностранны́ка (Inostranka) ‘A Foreign Woman’
Заповедник (Zapovednik) ‘The Reserve’
Компромисс (Kromiss) ‘The Compromise’
Литерату́ра про́должается (Literatura prodolžajaetsja) ‘Literature Carries On’
На́ши (Naši) ‘Ours’
Ремесло (Remeslo) ‘The Craft’
Соло́ на ундевуде́ (Solo na undervude) ‘Solo in Underwood’

Erofeev
Москва́ – Пету́шки (Moskva-Petuški) ‘Moscow – End of the Line’

Granin
Меся́ц вверх ногами́ (Mesjac vverx nogami) ‘The Moon Upside-down’
Grossman
   Все течет (Vse tečet) ‘Everything Flows’

Kazakov
   Запах хлеба (Zapax xleba) ‘The Smell of Bread’

Klimov
   Протоколы советских мудрецов (Protokoly sovetskix mudrecoy) ‘Protocols of the Soviet Sages’

Koneckij
   Невезучий Альфонс (Nevezučij Al’fons) ‘Luckless Alphonse’

Kozlov
   Ежик в тумане (Ežik v tumane) ‘Hedgehog in the Fog’

Žvanecskij
   Как это делается (Kak èto delaetcja) ‘How it is done’

Makanin
   Кавказский пленный (Kavkazskij plennyj) ‘Prisoner of the Caucasus’

Paustovskij
   Исаак Левitan (Isaak Levitan) ‘Isaak Levitan’
   Поток жизни (Potok žizni) ‘The Flow of Life’
   Правая рука (Pravaja ruka) ‘The Right Hand’
   Сказочник (Skazočnik) ‘The Story-Teller’
   Телеграмма (Telegramma) ‘Telegram’

Kukarkin
   Дельфинарий (Del’finarij) ‘Dolphinarium’
   Депутат (Deputat) ‘The Deputy’
   Любой ценой (Ljuboj cenoj) ‘At Any Price’
   Принц (Prince) ‘Prince’
   Метод беззакония (Metod bezzakonija) ‘Method of Lawlessness’
   Курсы повышения квалификации или слюнявая собачка Пати (Kursy povyšenija kvalifikacii ili sljunjavaja sobačka Pati)
     ‘Refresher Courses or The Slobbery Dog Pati’
   Долгая дорога в преисподнюю (Dolgaja doroga v preispodnjuju) ‘The Long Road to Hell’
   Уборщики ада (Uborščiki ada) ‘Cleaners of Hell’
Трудно быть героем (Trudno byt' geroem) ‘It’s Hard to be a Hero’
Бизерта - X (Bizerta - X) ‘Bizerta – X’
Я – Кукла (Ja – Kukla) ‘I am a Doll’
Трофейщик (Trofejščik) ‘Trofejiščik’
Сначала страх... (Snačala strach...) ‘First, Fear...’
Дальнобойщики (Dal'nobojščiki) ‘Truckers’
Дыхание кризиса (Dyxanie križisa) ‘The Breath Of Crisis’
Смерть всегда рядом (Smert’ vsegda rjadom) ‘Death is Always Close’
Первый (Pervyj) ‘The First’
Кунин
Интердевочка (Interdevočka) ‘International Girl’
Мандельштам, Н
Воспоминания (Vospominanija) ‘Memoirs’
Маринина
За все надо платить (Za vse nado platit’) ‘Everything must be paid for’
Игра на чужом поле (Igra na čužom pole) ‘Games on a Foreign Field’
Мужские игры (Mužskie igry) ‘Men’s Games’
Настя Каменская: (Nastja Kamenskaja:) ‘Nastja Kamenskaja:
Иллюзия греха (Illuzija grexa) ‘Illusion of Sin’
Не мешайте палачу (Ne mesajte palaču) ‘Don’t Disturb the Executioner’
Посмертный образ (Posmertnyj obraz) ‘Posthumous Manner’
Призрак музыки (Prizrak muziki) ‘The Ghost of the Music’
Светлый лик смерти (Svetlyj lik smerti) ‘The Bright Face of Death’
Седьмая жертва (Sed’aja žertva) ‘The Seventh Victim’
Смерть и немного любви (Smert’ i nemnogo ljubvi) ‘Death and a Little Love’
Стилист (Stilist) ‘Master of Style’
Шестерки умирают первыми (Sesterki umirajut pervymi) ‘Six Die First’
Чужая маска (Čužaja maska) ‘Strange Mask’
Черный список (Černyj spisok) ‘Black List’
Я умер вчера (Ja umer včera) ‘I Died Yesterday’
Украденный сон (Ukradennyyj son)  'Stolen Sleep'
Стечение обстоятельств (Stečenie obstojalet'stv)  'A Confluence of Circumstances'

Pelevin
Зомбификация (Zombificacijja)  'Zombification'
Имена олигархов на карте родины (Imena oligarxov na karte rodniny)  'Names of the Oligarchs on a Map of the Motherland'

Petruševskaja
Маленькая волшебница: кукольный роман (Malen'kaja volšebnica: kukol'nij roman)  'The Small Sorceress: a puppet novel'
Уроки музыки (Uroki muzyki)  'Music Lessons'
Сырая нога, или встреча друзей (Syraja noga, ili vstreča druzej)  'The Bare Leg, or a Meeting of Friends'
Чинzano (Činzano)  'Cinzano'
День рождения Смирновой (Den' roždenija Smirnovoj)  'Smirnova's Birthday'
Три девушки в голубом (Tri devuški v golubom)  'Three Girls in Blue'
Лестничная клетка (Lestničnaja kletka)  'The Stairwell'
Любовь (Ljubov')  'Love'
Анданте (Andante)  'Andante'
Квартира Коломбины (Kvartira Kolombiny)  'The Columbine's Flat'
Бабуля-Блюз (Babulja-Bljuz)  'The Granny-Bljuz'
Вставай, Анчутка! (Vstavaj, Ančutka!)  'Get up, Ančutka!'
Я болею за Швецию (Ja boleju za Šveiciju)  'I root for Sweden'
Стакан воды (Stakan vody)  'A Glass of Water'
Скамейка-премия (Skamejka-premija)  'Bench-Prize'
Дом и дерево (Dom i derevo)  'The House and the Tree'
Два окошка (Dva okoška)  'Two Windows'
Чемодан чепухи, или быстро хорошо не бывает (Čemodan čepuxi, ili bistro xorošo ne byvaet)  'A Trunk of Nonsense, or Nothing Good Happens Quickly'
Золотая богиня (Zolotaja boginja)  'The Golden Goddess'

Pikul
Богатство (Bogats'tvo)  'Wealth'
Фаворит (Favorit)  'The Favourite'
Васька Денисов, похититель свиней (Vaska Denisov, Kidnapper of Pigs)
Серафим (Serafim) 'Seraphim'
Выходной день (Vychodnoj den') 'A Day Off'
Домино (Domino) 'Domino'
Геркулес (Gerkules) 'Hercules'
Шоковая терапия (Sokovaja terapija) 'Shock Therapy'
Стланик (Stlanik) 'Elfin'
Красный Крест (Krasnyj Krest) 'Red Cross'
Заговор юристов (Zagovor juristov) 'The Lawyers' Plot'
Тифозный карантин (Tifoznyj karantin) 'Typhoid Quarantine'

Semenov
Аукцион (Aukcion) 'Auction'
Непримиримость (Neprimirimost') 'Irreconcilability'
Пресс-центр (Press-centr) 'Press Centre'
Семнадцать мгновений весны (Semanadcat' mgvenij vesny) 'Seventeen Moments of Spring'

Simonov
Живые и мертвые (Zhive i mertvye) 'The Living and the Dead'

Sokolov
Школа для дураков (Skola dlja durakov) 'School for Fools'
Между собакой и волком (Mezhdy sobakoj i volkom) 'Between a Dog and a Wolf'

Solouzin
Не жди у моря погody (Ne zdj u morja pobody) 'Do not wait for Fine Weather by the Sea'

Soljenicyn
Архипелаг ГУЛАГ (Arxeplag GULag) 'The Gulag Archipelago'

Strugackie
Далекая радуга (Dalekaja raduga) 'The Far Rainbow'
Гадкие лебеди (Gadkie lebedi) 'The Ugly Swans'
Град обреченный (Grad obrechnyyj) 'The Doomed City'
За миллиард лет до конца света (Za milliard let do konca sveta) 'Definitely Maybe'
Сукин

061.1 Tae Mbllli oc Tpoe

In Obitaemy ostrov 'The Inhabited Island'

Пикник на обочине (Piknik na obochine) 'Roadside Picnic'

Понедельник начинается в субботу (Ponedel'nik nachnatsja v subbotu) 'Monday Begins on Saturday'

Трудно быть богом (Trudno byt' bogom) 'It's Difficult to be a God'

Улитка на склоне (Ulitka na slope) 'Snail on the Slope'

Хищные вещи века (Xishnuye veschi veke) 'Predatory Things of the Century' (published in English as 'The Final Circle of Paradise')

Волны гасят ветер (Bolny gasjat veter) 'The Waves Extinguish the Wind'

Жук в муравейнике (Zuk v muravejnikhe) 'Beetle in the Anthill'

Христолюди (Xristoljudi) 'Christian People'

Šukšin

Мой зять украл машину дров (Moj zhat' ukral mašinu drov) 'My Son-in-Law Stole a Truckload of Wood!

Любавины (Ljubaviny) 'The Ljubavins'

ТарасенкоVa

Попутчики (Poputčiki) 'Fellow Travellers'

Терс

Что такое социалистический реализм? (Čto takoe socialističeski sh realizm?) 'What is socialist realism?'

Токорева

Система собак (Sistema sobak) 'System of the Dogs'

Толстая

Река Оккервиль (Reka Okkerwil) 'River Okkerwil'

Тритонов

Предварительные итоги (Predvaritel'nye itogi) 'Preliminary Conclusions'

Обмен (Obmen) 'Interchange'

Троепол'skij

Белый бим черное ухо (Belyj bim černoe yxo) 'White Beam, Black Ear'

Успенский

Эн-два-0 плюс Икс дважды (Èn-dva-0 plius Iks dvaždy) 'N-2-0 plus X twice'

Веллер

Байки скорой помощи (Bajki skoroy pomosći) 'Tales from an Ambulance'
Гуру (Guru) ‘The Guru’
Колечко (Kolečko) ‘The Ring’
Конь на один перегон (Kon' na odin peregon) ‘Horse for One Stage’
Котлетка (Kotletka) ‘Cutlet’
Легенды Невского проспекта (Legendy Nevskogo prospekta) ‘Legends of Nevskij Prospekt’
Зануда (Zanuda) ‘The Bore’
Ничего не происходит (Ničego ne proisxodit) ‘Nothing’s Happening’
Осуждение (Osuždenie) ‘Condemnation’
Приключения майора Звягина (Prikliučeniya majora Zvjagina) ‘The Adventures of Major Zvjagina’
Самовар (Samovar) ‘Samovar’
Травой поросло (Travoj poroslo) ‘Overgrown Grass’
Узкоколейка (Uzkokolejka) ‘The Narrow-Gauge Railway’

Вишневская
Галина (Galina) ‘Galina’

Войнович
Иванькида (Ivan'kiada) The Ivankiad
Жизнь и необычайные приключения солдата Ивана Чонкина (Žizn' i neobyčajnye priključeniya soldata Ivana Čonkina) The Life and Extraordinary Adventures of Private Ivan Chonkin
Москва 2042 (Moskva 2042) ‘Moscow 2042’

Зинов'ев
Иди на Голгофу (Idi na Golgofu) ‘Go to Golgotha’


