

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. ВИНОГРАДОВА

А. Я. ШАЙКЕВИЧ, В. М. АНДРЮЩЕНКО, Н. А. РЕБЕЦКАЯ

ДИСТРИБУТИВНО-
СТАТИСТИЧЕСКИЙ АНАЛИЗ
ЯЗЫКА РУССКОЙ ПРОЗЫ
1850—1870-х гг.

Том 1



ЯЗЫКИ СЛАВЯНСКОЙ КУЛЬТУРЫ
МОСКВА 2013

УДК 811.161.1
ББК 81.2 Рус
Ш 17

Издание подготовлено при финансовой поддержке
Российского гуманитарного научного фонда (РГНФ)
проект № 13-04-16009

Рецензенты:
д. ф. н. А. Ф. Журавлев, д. ф. н. С. А. Крылов

Шайкевич А. Я., Андриющенко В. М., Ребецкая Н. А.

Ш 17 Дистрибутивно-статистический анализ языка русской прозы 1850—1870-х гг. Т. 1. М.:
Языки славянской культуры, 2013. — 504 с. — (Studia philologica.)

ISBN 978-5-9551-0668-7

Цель дистрибутивно-статистического анализа состоит в открытии структуры языка на основе большого корпуса текстов. В настоящей трехтомной монографии этот формальный метод в полной мере прилагается к текстам русской прозы 1850—1870 гг. (около 15 млн словоупотреблений); а частично (в виде иллюстраций) к текстам на других языках.

Первый том включает три части:

Очерк развития метода;

Открытие регулярной морфологии в рамках графического слова;

Частотный словарь языка русской прозы 1850—1870 гг.

Первые две части адресованы лингвистам, особенно тем, кто интересуется лингвостатистикой. Частотный словарь будет интересен филологам-русистам. В существенно расширенном виде он представлен на компакт-диске.

ББК 81.2 Рус

Электронная версия данного издания является собственностью издательства,
и ее распространение без согласия издательства запрещается.

ISBN 978-5-9551-0668-7

© А. Я. Шайкевич, В. М. Андриющенко, Н. А. Ребецкая, 2013
© Языки славянской культуры, 2013

СОДЕРЖАНИЕ

Предисловие	5
Часть 1. Эволюция дистрибутивно-статистического анализа текстов.	7
1.1. Исторические предшественники	9
1.1.1. Таксономические проблемы в филологии и задачи ДСАТ.	9
1.1.2. Внешние влияния	12
1.2. Первые шаги на пути к формальному открытию системы языка по фактам речи	15
1.2.1. Статистико-комбинаторный метод Н. Д. Андреева	15
1.2.2. Изучение языка дешифровочными методами	20
1.2.3. Опыт изучения совместной встречаемости слов	23
1.3. Принципы работы ДСА	31
1.4. Черты поведения лингвистических элементов, используемые в ДСА	33
1.4.1. Синтагматическая сочетаемость	33
1.4.2. Статистические распределения лингвистических элементов	35
1.4.3. Позиционный анализ	41
1.4.4. Сходство и различие в окружениях лингвистических единиц	46
1.4.5. Совместная встречаемость лингвистических единиц	51
1.5. Интервалы текста в дистрибутивно-статистическом анализе.	52
1.5.1. Минимальный интервал	52
1.5.2. Средний интервал	71
1.5.3. Большой и максимальный интервалы	99
1.6. Лингвистические единицы и тексты	104
Часть 2. Дистрибутивно-статистический анализ в микроинтервале (статистическое открытие регулярной морфологии)	109
2.1. Позиции, n-граммы и парадигмы в статистическом открытии морфологии.	111
2.2. От протоаффиксов к парадигмам	115
2.2.1. Формирование протоаффиксов и протопарадигм	115
2.2.2. Формирование глагольных парадигм	117
2.2.3. Формирование адъективных парадигм	120
2.2.4. Формирование субстантивных парадигм	121
2.2.5. Уточнение парадигм	125
2.2.6. Префиксы	127
2.2.7. Парадигмы и размер исследуемого корпуса	129
2.3. Английский язык	131
2.4. Французский язык	137
2.5. Немецкий язык	141
2.6. Шведский язык	144
2.7. Итальянский язык	146
2.8. Латинский язык	148
2.9. Поиски парадигм — краткие итоги	153
Часть 3. Частотный словарь языка русской прозы 1850—1870-х гг.	157
Таблица 3.1	162
Таблица 3.2. Ранговый словарь 1 000 самых частых графических слов	481
Таблица 3.3. Ранговый словарь 1 500 самых частых лемм	486
Таблица 3.4. 100 самых частых существительных	493
Таблица 3.5. 100 самых частых прилагательных	494
Таблица 3.6. 100 самых частых глаголов	495
Литература	496