

The book cover features a blue and cyan color scheme. The background is a composite image showing a computer keyboard in the foreground, with a blurred background of a person's hands typing. Overlaid on this is a vertical column of white binary code (0s and 1s) on the left side. The title and author's name are printed in white on a dark blue, angular shape at the bottom.

**Рахимова Д.Р.**

**ПОСТ РЕДАКТИРОВАНИЕ  
КАЗАХСКОГО ЯЗЫКА  
В МАШИННОМ ПЕРЕВОДЕ**

ЦЕНТР ОПЕРАТИВНОЙ ПОЛИГРАФИИ

Д.Р. Рахимова

ПОСТ РЕДАКТИРОВАНИЕ  
КАЗАХСКОГО ЯЗЫКА В МАШИННОМ  
ПЕРЕВОДЕ

Алматы  
«Центр оперативной полиграфии»  
2022

УДК 80/81  
ББК 81.2  
Р 27

Рекомендовано к публикации Ученым советом  
РГП на ПХВ Института информационных  
и вычислительных технологий  
(протокол 8 от 26 сентября 2022г.)

**Рецензенты:**

д.т.н., профессор КазНУ им. аль-Фараби *У.А. Тулеев*  
д.э.н., профессор КазНУ им. аль-Фараби *К.Е. Кубаев*  
PhD, Заместитель генерального директора РГП  
«Института информационных и вычислительных  
технологий» *О.Ж. Мамырбаев*

**Рахимова Д.Р.**

Р 27 Пост редактирование казахского языка в машинном  
перевод: монография / Рахимова Д.Р. – Алматы: Центр  
оперативной полиграфии, 2022. – 150 с.  
ISBN 978-601-04-6054-6

Область машинного перевода в мировой науке является достаточно зрелой, разработаны многие формальные модели, алгоритмы, существуют достаточно высокого уровня системы машинного перевода. Однако, большинство результатов машинного перевода получены и применены к ведущим мировым языкам, таким как английский, французский, немецкий, русский, китайский. В приложениях к казахскому языку разработанных моделей и алгоритмов, или применений уже разработанных моделей и алгоритмов весьма мало. В данной книге представлены научные разработки по сбору и обработке лингвистических данных и моделей пост редактирования казахского языка в системе машинного перевода.

Монография представляет практический и теоретический интерес для научных работников, преподавателей, студентов, магистрантов и аспирантов специализирующихся в области информационных технологии, компьютерной лингвистики, а также смежных наук.

В монографии обобщаются результаты научных исследований, проведенных и профинансированных в рамках проекта МОН РК № АР08/052421 «Исследование и разработка системы постредактирования казахского языка в машинном переводе»

УДК 80/81  
ББК 81.2

ISBN 978-601-04-6054-6

© Рахимова Д.Р., 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	4
1 ИССЛЕДОВАНИЕ ПРОБЛЕМ ПОСТ РЕДАКТИРОВАНИЯ В МАШИННОМ ПЕРЕВОДЕ.....	6
1.1.1 Описание проблемы.....	6
1.1.2 Обзор научных исследований.....	9
1.2 Обзор видов и систем постредактирования в машинном переводе.....	15
1.3 Сравнительный анализ систем машинного перевода для казахского языка.....	26
1.4 Классификация ошибок при работе с системами машинного перевода.....	32
2 РАЗРАБОТКА МОДЕЛИ ОПРЕДЕЛЕНИЯ НЕИЗВЕСТНЫХ СЛОВ ПРИ ПОСТРЕДАКТИРОВАНИИ МАШИННОГО ПЕРЕВОДА КАЗАХСКОГО ЯЗЫКА.....	37
2.1 Обзор существующих методов.....	37
2.2 Описание модели определения неизвестных слов при постредактировании машинного перевода.....	37
2.3 Практические результаты.....	51
3 РАЗРАБОТКА СИНТЕТИЧЕСКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА ДЛЯ ОБУЧЕНИЯ СИСТЕМЫ ПОСТРЕДАКТИРОВАНИЯ МАШИННОГО ПЕРЕВОДА.....	53
3.1.1 Описание проблемы.....	53
3.1.2 Связанные работы.....	56
3.1.3 Разработка синтетического параллельного корпуса (TREEBANK).....	59
3.2 Задача выравнивания параллельного корпуса.....	66
3.2.1 Описание проблемы.....	66
3.2.2 Связанные работы.....	66
3.2.3 Технология выравнивания параллельного корпуса.....	68
3.2.4 Практические результаты и оценка.....	71
3.2.5 Технология метода выравнивания параллельного русско-казахского корпуса.....	75
3.2.6 Практические результаты и оценки.....	77
3.3 Разработка размеченного корпуса казахского языка.....	82
3.3.1 Обзор видов и систем размеченного корпуса.....	84
3.3.2 Разработка модели размеченного корпуса казахского языка.....	84
3.3.3 Программная реализация разработанного корпуса.....	88
3.4 Пополнение размеченного корпуса казахского языка.....	90
3.4.1 Разработка алгоритма автоматического пополнения текстов.....	92
3.4.2 Алгоритм сбора текстовых данных, поступающих в режиме реального времени.....	92
3.4.3 Индексирование документов с помощью признаков.....	98
3.4.4 Практические результаты.....	103
4 РАЗРАБОТКА POST-EDITING МОДЕЛИ ДЛЯ АНГЛО-КАЗАХСКОГО И РУССКО-КАЗАХСКОГО ПЕРЕВОДА.....	107
4.1 Архитектура модели и системы пост редактирования для англо-казахского и русско-казахского машинного перевода.....	109
4.2 Разработка метода анализа сложных предложения казахского языка.....	109
4.2.1 Описание проблемы.....	112
4.2.2 Связанные работы.....	112
4.2.3 Правила и алгоритмы определения и редактирования сложных предложений для англо-казахской пары языков.....	114
4.2.4 Практические результаты.....	116
4.2.5 Правила и алгоритмы определения и редактирования сложных предложений для русско-казахской пары языков.....	125
4.2.6 Практические результаты.....	129
4.3 Пост редактирование казахского языка на основе подхода машинного обучения.....	134
4.3.1 Языковые подходы с низким ресурсом.....	136
4.3.2 Архитектуры NMT.....	136
4.3.3 Результаты экспериментов.....	139
4.3.4 Оценка моделей.....	142
ЗАКЛЮЧЕНИЕ.....	145
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	146
	148

Научное издание

Рахимова Диана Рамазановна

**ПОСТ РЕДАКТИРОВАНИЕ КАЗАХСКОГО ЯЗЫКА В  
МАШИННОМ ПЕРЕВОДЕ**

Монография

Выпускающий редактор Д.А. Балгабаев

ИБ №11694

Формат 60x84<sup>1/16</sup>. Бумага офсетная. Печать цифровая.

Объем 10, 25 п.л. Тираж 500 экз. Заказ № 118

Издательский дом «Центр оперативной полиграфии»  
г. Алматы, Масанчи 23.