

К Ад 10.02.06

З - 14

А - 444

МИНИСТЕРСТВО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ЧУВАШСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. И. Н. УЛЬЯНОВА



На правах рукописи
УДК 809.434: 800.853

Зайцева Вера Петровна

ИССЛЕДОВАНИЕ ЧАСТОТНОСТИ
УПОТРЕБЛЕНИЯ СЛОВ В РАЗЛИЧНЫХ ТИПАХ
ЧУВАШСКИХ ТЕКСТОВ

Специальность 10.02.06 - тюркские языки

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата филологических наук

Чебоксары

2000

15
Работа выполнена на кафедре чувашского языка и литературы
Чувашского государственного педагогического
университета им. И. Я. Яковлева

КАД @444

Научные

10.02.06

3 17 Зайцева, В. П.

Исследование частотности
употребления слов в

различных типах чувашских

Общественные

ВОЗВРАТИТЕ КНИГУ НЕ ПОЗЖЕ

обозначенного здесь срока

ин
са-

М. т. 3 50 1990 г. т. 5000

Автореферат разослан 27 июня 2000 г.

Ученый секретарь
диссертационного совета,
кандидат филологических наук

А-444
ЧУВАШСКАЯ РЕСПУБЛИКА


Ю. М. Виноградов

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В настоящее время использование компьютерных технологий стало практически необходимым для реализации самых разных исследовательских задач. Компьютерная поддержка развития и функционирования чувашского языка во всех сферах общественной жизни и структуры управления в состоянии сыграть роль катализатора и фактора необходимости в расширении его функциональной роли.

Актуальность работы. В исследовании языковых явлений в зарубежном и отечественном языкознании накоплен большой опыт применения количественных методов с помощью компьютерных технологий. Статистические методы к большому объему текстов с использованием автоматизации в чувашском языкознании почти не применялись. Актуальность данной проблемы определяется необходимостью создания с учетом принципа частотного распределения лексических единиц языка учебных словарей, школьных и вузовских учебников, методических пособий. Полученная база может стать частью машинного фонда чувашского языка и использоваться для его дальнейшего исследования.

Цель работы. Целью настоящей работы является выявление типологических особенностей чувашского языка с помощью использования статистических методов и современных компьютерных технологий, анализ типологической и стилиевой структуры чувашского текста и проведение сравнения с другими тюркскими языками. Цель работы состоит в создании экспериментального образца базы данных (как подфона машинного фонда чувашского языка) для лингвистических исследований и ее апробации.

Эта общая цель реализуется путем выполнения следующих конкретных задач:

- 1) определения объема репрезентативной выборки чувашских текстов;
 - 2) составления частотных словарей для отдельных видов чувашских текстов (в частности, для подъязыка публицистики, прозы, учебно-методической литературы);
 - 3) выявления таких статистических характеристик чувашских текстов, как средняя длина предложения (в словах), средняя длина слова (в буквах), максимальная и минимальная длина предложения, максимальная и минимальная длина слова;
 - 4) выявления типологических особенностей чувашских текстов по отношению к другим тюркским языкам на уровне лексики;
 - 5) сравнения лексико-стилевых особенностей в различных типах текстов;
 - 6) изучения общеупотребительной лексики, характерной для всех типов текстов (публицистических, художественных, учебно-методических).
- Научная новизна нашего исследования заключается в том, чтобы с помощью статистических методов применительно к большому объему чувашских текстов и применения компьютерных технологий получить общую картину лексико-статистической структуры публицистических,

художественных, учебно-методических текстов чувашского языка.

В исследовании разработаны конкретные подходы и методы лингвостатистического анализа чувашских текстов, определены и обоснованы критерии отбора текстов для лингвостатистического анализа чувашских текстов, определены возможности автоматизации многих процессов, связанных с обработкой чувашских текстов.

Теоретическое и практическое значение работы. В теоретическом аспекте результаты настоящей работы могут использоваться при теоретико-методической разработке лексикологии, лексикографии и морфологии чувашского языка, в типологических исследованиях тюркских языков в целом и чувашского языка в частности.

Практическое применение результатов исследования возможно:

- 1) в сравнительно-сопоставительном изучении характеристик тюркских языков;
- 2) при лексикографическом и грамматическом нормировании чувашского языка, в частности, в отраслевой и учебной лексикографии;
- 3) в методике обучения и преподавания чувашского языка;
- 4) при создании систем автоматизированной переработки текстовой информации.

Практическая ценность работы определяется тем, что использованные тексты для исследования и полученные словари составляют текстовую и словарную базы данных и закладываются как основа машинного фонда чувашского языка. Результаты проведенных исследований могут быть использованы для изучения различных языковых аспектов, при подготовке к изданию словарей и учебных материалов по чувашскому языку.

Материал и методика работы. Исследование проведено на материале современных газетных, учебно-методических текстов и художественных произведений некоторых авторов. Были исследованы газетные тексты "Хыпар" (1999 г.) объемом 220 190 словоупотреблений, тексты книг для внеклассного чтения "Шуҫӑм", "Сесӑл", "Сӑлкуҫ", объем которых составляет 53 161 словоупотребление, произведения чувашских писателей Юхмы М.Н. "Хӗвел хапхи" (54 666 словоупотреблений) и Тимофеева И.Д. (Вутлан) "Ытла та вӑрттан юрату" (75 867 словоупотреблений). Общий объем выборки составлял 403 884 словоупотребления. Для решения поставленных задач на различных этапах использовались компьютерная обработка чувашских материалов, методы квантитативно-лингвистического исследования, лексико-грамматический анализ текстов.

В результате обработки этих текстов на FoxPro получены 3 вида частотных словарей:

- 1) ранговый частотный словарь;
- 2) алфавитно-частотный словарь;
- 3) обратный частотный словарь.

Эти словари послужили для дальнейшего лингвистического анализа.

Апробация работы и публикации. Основные положения и результаты исследования были сообщены на научно-практических конференциях аспирантов и докторантов ЧГУ им. И.Я.Яковлева (1998-2000), на

конференции “Проблемы информатизации образования в Чувашской Республике” (Чебоксары, 1997), на всероссийской научно-практической конференции “Проблемы изучения и преподавания филологических наук” (Стерлитамак, 1999), на международной научно-практической конференции, посвященной 80-летию Чувашской Республики (Чебоксары, 2000).

Структура диссертации. Работа состоит из введения, основной части, содержащей три главы, заключения, списка использованных источников и литературы, приложения, где представлены частотные словари исследуемых текстов.

II. ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении освещаются актуальность, цели и задачи исследования, обосновываются актуальность и новизна работы, определяется ее теоретическая и практическая значимость.

Первая глава “Перспективы развития чувашского языкознания в современных условиях” состоит из трех параграфов.

В первом параграфе раскрываются основные тенденции использования информационных технологий в современном чувашском языке.

В настоящее время информационные технологии становятся неотъемлемыми атрибутами современной жизни. Данные технологии используются практически во всех областях человеческой деятельности, в том числе и в языкознании. Включение национального языка в информационные технологии помогает достижению многих целей. Возрождение функционального значения чувашского языка как хранителя этнических и культурных традиций народа не может сегодня решаться методами 30-х годов. Радикально изменились условия жизни, условия функционирования языка. Одной из первоочередных, судьбоносных задач для чувашского народа является задача компьютеризация чувашского языка.

“Компьютеризация” чувашского языка как функционального и государственного будет способствовать укреплению его роли и значения. Последовательная политика компьютерной поддержки и использования чувашского языка во всех сферах государственной, хозяйственной и культурной жизни должна сыграть роль катализатора в расширении его функционирования. В итоге компьютерный подход к проблемам поддержки языка позволит восстановить былое богатство национальной культуры и национального языка, вернуть ему положение связующего и цементирующего элемента нации.

Грамотное решение задач компьютерной поддержки чувашского языка как государственного требует концептуального подхода, продуманного целостного, перспективного планирования и постоянной координации усилий научных и практических профессиональных коллективов разного профиля. Финансирование работ по проблеме должно производиться с ориентацией на создание республиканского центра по проблемам компьютерного обеспечения национального языка и культуры.

Внедрение компьютерных технологий в чувашский язык приведет к следующим положительным результатам:

- повысится эффективность и оперативность внедрения чувашского языка как государственного;
- возникнут благоприятные условия для более успешного овладения языком и его широкого распространения;
- повысится эффективность работы специалистов в издательской деятельности, сфере образования, делопроизводстве;
- возникнут условия для формирования благоприятного общественного сознания в области компьютерной технологии, в компьютерном образовании населения;
- возникнут предпосылки для восстановления словарного фонда и самобытности чувашского языка, поднимется престиж Чувашской Республики.

Главной сложностью решения задач является отсутствие научных исследований и научно-практических работ в области функционирования чувашского языка в компьютерных технологиях. Поэтому при реализации программы внедрения чувашского языка в информационные технологии важным является также и научный подход в изучении особенностей чувашского языка, который во многом может облегчить решение проблем компьютеризации.

Во втором параграфе “Статистические методы и их применение в лингвистике” обосновывается использование статистических методов, даются сведения об истории, типах, составлении частотных словарей и об их применении. На основе обзора литературы по теме исследования выделяются основные сферы применения статистических методов в языкознании, в том числе при составлении частотных словарей, прослеживается история их создания, рассматриваются типы, методика создания, применение таких видов словарей.

Статистические методы дают возможность глубже проникнуть в законы языка и речи, с их помощью можно определить некоторые закономерности, которые не обнаруживаются обычными лингвистическими методами. Эти же методы могут использоваться для изучения особенностей функциональных стилей языка, исторических и типологических исследований языка, машинного перевода и автоматического реферирования текста, успешно применяться в исследовании своеобразия стиля того или иного писателя, в сравнительно-исторических исследованиях языка, таких как историческое развитие словаря, история грамматического строя. Данные по статистическому исследованию речи могут использоваться при автоматическом анализе текстов, что тесно связано с машинным переводом, требующим принципиально нового подхода к исследованию речи.

В третьем параграфе “Использование информационных технологий в лингвистических исследованиях” дается краткое представление о современных информационных технологиях, определены компьютерной базы данных как взаимосвязанных, определенным образом структурированных данных, которые можно обрабатывать различными алгоритмическими программами. Констатируется, что компьютерное представление информации в виде баз данных в настоящее время широко используется в самых разных областях языкознания. Приводятся примеры разнообразных лингвистических

данных, разрабатываемых в России, Европе и США.

Накопленный опыт использования ЭВМ в лингвистических исследованиях и лексикографии свидетельствует, что ЭВМ постепенно становится необходимым и привычным инструментом лингвиста и лексикографа, помогающим выполнять самую трудоемкую и рутинную работу по составлению частотных словарей, сортировке, статистической обработке материала. Основой автоматизации лингвистических исследований являются текстовые и словарные базы данных. Труд, затрачиваемый на их формирование, быстро окупается благодаря возможности их многократного и многоцелевого использования. Анализ зарубежного опыта автоматизации лингвистических исследований показал, что ЭВМ является самой удобной формой хранения текстов и их обработки.

В нашей стране автоматизация лингвистических исследований и лексикографических работ в настоящее время развивается в рамках создания машинного фонда русского языка и машинных фондов языков народов бывшего СССР.

Первопричина создания машинного фонда чувашского языка (МФ ЧЯ) заключается в том, что в эпоху информатизации и возрастания роли вычислительной техники и автоматизации в жизни общества роль естественного языка многократно возрастает. Создание МФ ЧЯ активизирует процессы внедрения компьютерных технологий в республике, ускоряет вхождение Чувашской Республики в мировое информационное сообщество.

Создание МФ позволит значительно активизировать научные изыскания, особенно в области гуманитарных наук, будет способствовать расширению фронта исследований, углублению их задач и возрастанию оперативности и точности получения результатов. Накопление и сохранение всей национально-значимой информации на машинных носителях, создание условий для эффективного научного анализа культурного и исторического наследия нации с использованием современных информационных методологий в состоянии оказать заметное влияние на темпы и результативность исследований в этих областях. Политическое и общекультурное значение создания МФ ЧЯ заключается в том, что обмен данными МФ поднимет авторитет и расширит функции чувашского языка как государственного и межгосударственного общения; увеличение номенклатуры и улучшение качества изданий словарей и пособий по чувашскому языку будет способствовать большей информированности мировой общественности о чувашах и чувашском языке.

МФ ЧЯ в первую очередь должен включить следующие базы данных как подфонды:

1. Генеральный словарь ЧЯ – инвентарь всех слов ЧЯ, снабженный необходимой информацией, дающих представление о морфологических, лексических, грамматических, орфографических, орфоэпических, словообразовательных и другим особенностях каждого отдельного слова.

2. Иллюстративно-текстовый фонд ЧЯ.

3. Терминологический фонд ЧЯ.

4. Морфемный фонд ЧЯ.

5. Фонетический фонд ЧЯ.

6. Историко-лингвистический фонд.
7. Фразеологический фонд.
8. Диалектологический фонд.

Кроме перечисленных пунктов МФ ЧЯ должен содержать подфонды прикладного характера, т.е. частотные словари различных видов. Создание частотных словарей связано с тем, что уровень информатизации в Чувашской Республике очень высок, а компьютерная лингвистика чувашского языка совершенно не переработана; очень высока потребность именно в частотных словарях, т.к. их отсутствие сдерживает оперативный выпуск качественной учебной и научно-методической литературы.

Вторая глава "Выбор материала и методика исследования" состоит из 4-х параграфов.

В первом параграфе дается описание построения количественной модели для исследования, построение которой осуществляется в следующие четыре этапа: 1) определение тематической структуры и объема исследуемого текста и подбор реальных текстов; 2) получение ранговых частотных, алфавитно-частотных и обратных частотных словарей; 3) определение основных результативных статистических характеристик в различных чувашских текстах; 4) сравнение количественных характеристик с аналогичными данными других языков.

В задачу нашего исследования входило выявление количественных и качественных критериев отбора текстов для статистического изучения чувашского языка. Параграф "Определение тематической структуры корпуса" представляет собой обоснование выбора данных видов текстов: публицистических, художественных, лексики учебной детской литературы.

Выбор подязыка публицистических текстов объясняется следующими особенностями:

1. В результате особенностей своего функционирования язык газеты в гораздо большей степени, чем другие разновидности общенародного языка, отражает влияние экстралингвистических и, в частности, социальных факторов.

2. Обладая широкими коммуникативными возможностями, газета фиксирует все новые лингвистические единицы, употребляемые в литературной и в разговорной речи.

3. Поскольку газетным текстам присущи периодичность, массовость, широкое распространение среди населения, язык газеты способствует закреплению языковой нормы в устной и письменной речи.

4. Пресса, наряду с радио и телевизионными передачами, является средством коммуникации и характерна тем, что имеет "массового читателя" и "массового писателя", а последнее устраняет влияние манеры изложения материала одним автором.

5. Язык газетных текстов дает возможность глубже проникнуть в процессы, происходящие в литературном языке вообще. Газетный язык оказывает большое влияние на устную речь, а также на язык художественной литературы.

Получение частотного словаря подязыка чувашской публицистики позволит наглядно увидеть насколько часто встречаются неологизмы —

полукальки в нашей обычной речи. Тогда, когда "лингвисты, понимая сущность и закономерности развития языка, в контакте с журналистами из редакций газет, журналов, радио и телевидения, с учетом всех сторон дела, возьмутся за словотворную работу"¹, чувашский язык будет прогрессивно развиваться и совершенствоваться.

Изучение языка и стиля художественных произведений является одним из актуальных вопросов современного языкознания. Речевая культура и само мастерство писателя не могут быть поняты, если не будут полно и глубоко исследованы те процессы, которые происходят в языке художественных произведений.

В зависимости от того, какая сторона, какой объект языка интересует исследователя, используются те или иные приемы и методы наблюдения. При этом важным является выявление индивидуальных особенностей словарного состава произведений, его лексики, которая выступает как главное орудие в достижении максимальной выразительности описываемого. Изучение детской речи как национальной формы общения, приобщающей человека к окружающей среде, должно определяться прежде всего тем, что овладение устной, а затем и письменной речью в детском возрасте является всеобщим и самым естественным способом приобщения формирующегося человека и всего поколения к своему национальному коллективу, объединенному собственным языком.²

Важное место в этом процессе выполняет освоение младшими школьниками лексики и семантики учебных и литературных текстов. Они являются для ребенка не только важнейшим источником новой информации о внешнем мире, но и выполняют основную нормализующую функцию при овладении ребенком литературным языком.

В качестве конкретной задачи работы становится задача изучения количественными методами лексики чувашских детских текстов, выступающих в качестве основных источников развития и нормализации родного языка у младших школьников-чувашей. Основным методическим приемом, позволяющим выполнить эту задачу, является составление частотного словаря по указанным выше текстам.

Во третьем параграфе "Определение достаточного объема выборки" описывается определение достаточного объема выборки для лингвостатистического исследования чувашских текстов. При определении объема выборки мы придерживались традиционной методики.

Используя формулу для вычисления и постоянные величины, которые выработаны в лингвостатистике, мы определили необходимый объем выборки для статистического исследования чувашских текстов. Определив нужные величины, мы выяснили, что для составления частотного словаря чувашского языка достаточна выборка объемом в 400 тыс. словоупотреблений.

1. Федотов М.Р. Чувашский язык. Отношение к алтайским и финно-угорским языкам. Историческая грамматика. - Чебоксары: ЧГУ, 1996. - С.306.

2. Гвоздев А.Н. Вопросы изучения детской речи. - М.: Изд-во АПН РСФСР, 1961. - С.9.

В четвертом параграфе “Используемое программное обеспечение” раскрываются возможности и основные характеристики компьютерных программ “Помощник филолога”.

Третья глава “Информационные и лингвистические особенности чувашских текстов” посвящена лингвостатистическому изучению наиболее частотной лексики различных типов текстов, проводится обоснование статистических характеристик, получаемых в результате количественного исследования отдельных типов текстов.

Были изучены такие статистические характеристики чувашского языка, как средняя частота (повторяемость словоформ), покрываемость и заполняемость чувашских текстов, информационные оценки чувашской словоформы, средняя длина словоформы в буквах для разных типов текстов.

При делении общего объема выборки N на общее количество разных словоформ L полученная величина \tilde{F} будет характеризовать то количество текстовых словоупотреблений, которое в среднем приходится на одну словарную лексическую единицу. Чем выше значение \tilde{F} , тем меньше разных словоформ в исследуемом тексте; чем ниже, тем больше словоформ или разнообразнее словарь текста. Коэффициент разнообразия словаря C определяется соотношением количества разных словоформ к числу всех словоупотреблений.

Таким образом:

$$\tilde{F} = \frac{N}{L} - \text{средняя повторяемость словоформ;}$$

$$C = \frac{L}{N} - \text{коэффициент разнообразия словаря.}$$

где

N - объем текста;

L - объем словаря.

В ходе нашего исследования были получены следующие показатели. Из газетных текстов общим объемом 220 190 СУ был получен частотный словарь объемом 32 460 лексем. Для других типов текстов эти показатели равны:

- 1) для произведения Юхма М.Н. “Хёвел хапхи” $N=75\ 867$ СУ, $L=12\ 185$ СФ;
- 2) для произведения Тимофеева И.Д. “Ытла та вартган юрату” $N=54\ 666$ СУ, $L=7\ 916$ СФ;
- 3) для детских текстов $N=53161$ СУ, $L=7\ 829$ СФ.

Таким образом средняя повторяемость для публицистических текстов, для произведений “Хёвел хапхи”, “Ытла та вартган юрату”, детских текстов соответственно равна 6,78; 6,22; 6,90; 6,79. Был вычислен коэффициент

лексического разнообразия частотного словаря, который для публицистических текстов оказался равным 0,146. Для произведений "Хёвел хапхи" и "Ытла та варттан юрату" этот показатель оказался равным соответственно 0,160 и 0,144. Для детских текстов $C=0,147$. Сравнивая \bar{F} произведений "Хёвел хапхи" и "Ытла та варттан юрату" можно сказать,

что лексическое разнообразие первого произведения оказалось богаче лексического разнообразия второго произведения, т.е. Юхма М.И. в своем произведении использует больше различных словоформ (Сравнение по одному произведению каждого писателя не свидетельствует о том, что язык одного писателя богаче языка другого писателя; для других произведений этих же писателей картина лексического разнообразия словаря может оказаться обратной). Лексическое разнообразие частотных словарей публицистических и детских текстов оказалось примерно равной.

В публицистических текстах наиболее частые лексические словоформы (до абсолютной частоты 9) составляют 3384 словоформ, что составляет 10,4 % от общего количества словоупотреблений. Словоформы с абсолютной частотой 1 покрывают 366 лексических единиц, т.е. 49,3 % газетных текстов выбранного объема. Отсюда можно сделать вывод, что половину газетных текстов покрывают редко встречающиеся, мало употребляемые словоформы. Покрываемость с частотой 1, т.е. редко и мало употребляемых словоформ в произведениях "Хёвел хапхи", "Ытла та варттан юрату" одинакова и составляет 51 %. В детских текстах данная покрываемость чуть больше и равна 55 %. Покрываемость словоформ в произведениях "Хёвел хапхи", "Ытла та варттан юрату" и в детских текстах с частотой больше 9 составляет соответственно 10,8 %, 12,9 %, 6,48 %. Как видим, в детских текстах, по сравнению с другими типами, покрываемость словоформами с абсолютной частотой больше 9 почти в два раза меньше.

Увеличение объема выборки для чувашского языка не дает того быстрого эффекта повышения средней повторяемости отдельных словоформ и роста устойчивости частот, которые мы наблюдаем во флективно-аналитических языках, например в английском, румынском (молдавском) и других западноевропейских языках.

Также были исследованы словоформы, общие для всех типов текстов, и специфичная лексика, характерная для каждого типа текстов.

В результате лингвостатистического анализа общей лексики, характерной для данных типов текстов, было установлено:

- Самую активную часть общей лексики составляют глаголы (37,9 %). В десятку самых активных глаголов входят следующие словоформы: *терё* (307), *терсё* (209), *ячё* (212), *пулчё* (211), *тытӑнчё* (201), *пулать* (199), *кайрё* (183), *илчё* (166), *пырать* (151), *тетпёр* (108). В результате статистического изучения общей лексики установлено, что спрягаемая форма глагола чаще встречается в прошедшем времени (58%), 30% встречается в настоящем времени и в будущем времени - 12%. Наибольшую активность проявляют неспрягаемые формы глагола - причастие, деспричастие, инфинитив, которые составляют 35,8% глаголов. Причастия и деспричастия по частоте употребления почти одинаковы (14,32% и 14,36 %). Среди

причастий в общей лексике преобладают причастия прошедшего времени (74%). К наиболее часто встречающимся причастиям такой формы можно отнести *тепӕ* (172), *пурӕппӕ* (100). Причастия долженствования, например, *пудмалла* (177), *каймалла* (51) встречаются почти в 7-8 раз меньше. Причастия будущего времени употребляются еще меньше. К ним можно отнести причастия *дӕлес* (31), *кайс* (19). Причастия настоящего времени составляют наименьшую часть. В первую десятку наиболее часто встречающихся причастий входят *пудмалла* (177), *тепӕ* (172), *илӕ* (144), *тухнӕ* (103), *пурӕппӕ* (100), *япӕ* (96), *калавӕ* (90), *тытӕннӕ* (80), *пудмап* (76), *кӕпӕ* (73). В число наиболее часто встречающихся деепричастий входят *тесен* (421), *илсе* (257), *пудсап* (218), *пудса* (207), *каласан* (166), *туса* (151), *тухса* (145), *тытса* (90), *пырса* (86), *пудласа* (70).

- В зоне наиболее употребительной лексики находятся имена существительные. Эта часть речи покрывает 21,4% словоформ общей лексики. Но эти части речи характеризуются меньшей повторяемостью. Во-первых, это объясняется тем, что чувашская лексика пополняется новыми словами, приходящими из других языков. Во-вторых, это объясняется тем, что часть имен существительных заменяется местоимениями. В десятку самых активных существительных входят следующие слова: *чӕван* (544) - русск. чуваш, *халӕх* (272) - русск. народ, *ӕс* (184) - русск. работа, *сып* (142) - русск. человек, *хӕвл* (140) - русск. солнце, *шыв* (100) - русск. вода, *сӕмах* (99) - русск. слово, *пус* (95) - русск. голова, *хыпар* (80) - русск. новость, *тӕнче* (67) - земля, мир. Нужно отметить, что в последующих чувашских текстах чаще употребляются имена существительные в единственном числе. Они составляют 85,96 % всех существительных. Если сравнить применение существительных в притяжательной форме и в форме определения, то форма определения почти не применяется. Выделительные формы имен существительных в форме определения, которые образованы аффиксом *-я* встречаются только в 0,2% из них. Существительные в форме определения, образованные аффиксом *-скер*, в частотном словаре общей лексики, образованном с помощью логических операций пересечений частотных словарей публицистических, художественных и детских текстов, не встречаются. 36,13 % существительных принимают притяжательную форму, из которых 90,9% являются формой 3-го лица. Существительные 1-го и 2-го лица применяются в чувашских текстах очень мало. Можно привести такие примеры, как *хӕрӕм* (36), *чӕлхӕс* (3). Наил исследования показали, что в общей лексике существительные дательного падежа не применяются. Очень редко употребляются существительные притяжательного падежа. Они составляют 1,45% всех существительных. К ним можно отнести такие слова, как *сӕршывӕн* (9), *сыппӕн* (11), *халӕхӕн* (18). Существительные причинно-целевого падежа встречаются еще в два раза меньше, чем существительные притяжательного падежа. Например, к таким существительным можно отнести словоформы *халӕхтӕн* (10), *кушӕн* (6). Существительные именительного падежа составляют 32,1%, местного падежа - 12%, исходного - 4%, творительного - 9,4%. Существительные дательного падежа покрывают 36% всех существительных.

В ходе анализа местоимений выяснилось, что наиболее употребительными местоимениями являются *ана* (647), *пирён* (394), *сак* (385), *вёсем* (347), *эптр* (340), *эсё* (322), *уя́н* (291), *сапа* (227), *хэйён* (205), *эсвр* (200). Они составляют 6,7% общей лексики. Высокочастотными являются личные местоимения, которые покрывают 32% всех местоимений. 18% местоимений покрывают возвратные местоимения. Неопределённые местоимения употребляются меньше (14%). 12%, 10%, 7% составляют вопросительные, отрицательные, определённые местоимения.

Примерно столько же по частоте употребления встречаются прилагательные (6,3%). 70,5% из этих прилагательных являются качественными. К наиболее употребительным прилагательным можно отнести *тесляйё* (147), *асла́*(191), *лайа́х* (172), *чапла́* (113), *сёпё* (112), *хитре* (112), *пысак* (106), *самра́к* (89), *тёрлё* (81), *шурá* (82). 6,33% текстов общей лексики покрывают наречия. В ходе анализа нами были проанализированы служебные части речи (союзы, частицы, послелоги). Послелоги составляют 2,6%, союзы и частицы - 5%. Распределение конкретных послелогов по частоте употребления выглядит таким образом: *синё* (467), *синчён* (395), *патис* (280), *хысёя́н* (173), *валли* (144), *хушпя́нче* (96), *патёпче* (66).

Столкнувшись с тем, что определение части речи без окружения данной словоформы в данном тексте невозможно и познакомившись с опытом других исследователей, мы решили, что все словоформы, которые могут являться разными частями речи, целесообразно отнести в отдельную группу.

Нашими исследованиями установлено, что средняя длина словоформы для каждого типа текстов может быть разной. Для публицистических текстов она равна 8,16, для художественных текстов - 7,78, для детских текстов - 7,33. Отсюда можно сделать вывод, что средняя длина словоформы в какой-то степени может служить показателем для определения типа текстов.

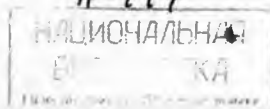
Словоформы, длина (в буквах) которых больше 20, встречаются в основном в газетных текстах. Обычно такими словами являются парные слова. В качестве примера можно привести такие словоформы: *предприя́ти-орга́низаци́еспче*, *предприя́ти-учрежде́нсен*, *ча́та́мла́хё-ту́сё́млё́хёнчсн*.

Для публицистических текстов характерна большая частотность для таких букв, как *о, с, в, г, д, с, н, й, м, о, у, ф, ы, ь, э*. В общеупотребительной лексике чаще применяются такие буквы, как *а, ä, ё, и, с, с, х, ч, ш, ы, ь*.

Заключение содержит основные результаты и краткие выводы исследования:

1. Использование информационных технологий в современном чувашском языке приведет к следующим положительным результатам:

- повысится эффективность и оперативность внедрения чувашского языка как государственного;
- возникнут благоприятные условия для более успешного овладения языком и его широкого распространения;
- повысится эффективность работы специалистов в издательской деятельности, сфере образования;



возникнут предпосылки для восстановления словарного фонда и самобытности чувашского языка;

· возрастает эффективность обмена и использования информационных технологий всех сферах деятельности, поднимется престиж Чувашской Республики.

2. Частотный словарь является одним из результатов применения статистических методов. Существующие частотные словари классифицируются в зависимости от расположения словарного материала, объема, выборки, содержания и формы текстов, входных единиц, их численных характеристик и техники составления словаря.

3. Частотные словари применяются в различных областях – в традиционной лексикографии, в теориях речевой деятельности, в лингвистической типологии, в методике обучения языку, в инженерной лингвистике, в психолингвистике.

4. Накопленный опыт использования ЭВМ в лингвистических исследованиях свидетельствует о том, что ЭВМ помогает выполнять самую трудоемкую и рутинную работу по статистической обработке материала, сортировке и т.д.

5. Основой автоматизации лингвистических исследований и лексикографии являются текстовые и словарные базы данных.

6. Выбор текстов или словарей в качестве основного источника национального машинного фонда определяется общей концепцией разработчиков.

7. Создание текстовой и словарной баз данных чувашского языка закладывает основу машинного фонда чувашского языка.

8. Автоматическая обработка базы данных дала возможность получить частотность букв, ранговые частотные, алфавитно-частотные, обратные частотные словари и интересный статистический материал, уточняющий некоторые характеристики чувашских текстов: средняя частота словоформы, покрываемость и заполняемость чувашских текстов, информационные оценки чувашской словоформы.

9. В результате лингвостатистического анализа общей лексики, характерной для всех типов текстов, выяснилось, что в ее состав входят собственные чувашские слова и заимствования из русского языка. Наиболее частотную зону словаря общей лексики составляют глаголы, существительные и местоимения. Наименьшей покрываемостью обладают прилагательные и наречия. Общая лексика характеризуется активностью существительных единственного числа. Активность проявляют существительные в приглагольной форме III лица. В результате статистического анализа надежной структуры выяснилось, что наиболее распространенной формой является дательный. Наибольшую активность проявляют глаголы прошедшего времени.

10. Наш опыт показывает, что даже на текстовой и словарной базах данных сравнительно небольшого объема могут быть получены результаты, которые находят широкое применение как в лингвистических исследованиях, так и в практической деятельности по подготовке учебников и словарей.

Однако, ценность текстовой и словарной баз данных будет расти с увеличением их объема, т.е. при продолжении работы в этом направлении.

11. Необходимо организовать работу таким образом, чтобы результаты, полученные на текстовой и словарной базах данных, сразу же могли быть использованы и машинный фонд чувашского языка превратился бы в естественную составляющую исследовательского процесса.

По теме диссертации опубликованы следующие работы:

1. Зайцева В.П. Постановка лингвистических задач с помощью информационных технологий // Сборник научных трудов студентов, аспирантов и докторантов. - Чебоксары: ЧГПУ им. И.Я.Яковлева, 1998. - Выпуск III. - С.12.
2. Зайцева В.П. Частотные словари и к методике их составления // Елизавета Федоровна Васильева: к 65-летию со дня рождения. - Чебоксары. - С.49-52.
3. Зайцева В.П., Ванюлин А.Н. Статистический анализ чувашских текстов // Сборник научных трудов студентов, аспирантов и докторантов. - Чебоксары: ЧГПУ им. И.Я.Яковлева, 1999. - Выпуск V. С. 56-60.
4. Зайцева В.П. Некоторые статистические характеристики чувашского языка (на материале публицистического стиля) // Проблемы изучения и преподавания филологических наук. Сборник материалов. - Часть IV. - (Татарская и чувашская филология) - Стерлитамак, 1999. - С.151-153.
5. Зайцева В.П. Частотный словарь и его применение при обучении языку // Чувашский язык и литература: теория и методика. Сборник статей на чуваш. и рус.яз. - Чебоксары: Чуваш. гос. ин-т гуманит.наук, 1999. - С.35-40.
6. Зайцева В.П. К концепции внедрения чувашского языка в компьютерные технологии // Сборник научных трудов докторантов, научных сотрудников и аспирантов. Выпуск VI. - С.144-147.
7. Зайцева В.П. База данных как основа для лингвистических исследований (на материале чувашского языка) // Чувашская республика на рубеже тысячелетий: история, экономика, культура. Тезисы международной научно-практической конференции, посвященной 80-летию Чувашской Республики, 22 июня 2000 г. Чебоксары, 2000. - С.315-318.
8. Зайцева В.П. К концепции создания машинного фонда чувашского языка. (В производстве).