

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  
ВЫСШЕГО ОБРАЗОВАНИЯ  
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

**Методические указания**  
**по выполнению практических работ по дисциплине**  
**«Автоматизированная обработка и перевод текста»**

Направление подготовки	45.04.02 Лингвистика
Направленность (профиль)	Современные методы прикладной лингвистики и перевода
Год начала обучения	2026
Форма обучения	очная
Реализуется в семестре	2

**Ставрополь**  
**2026**

## **Введение**

Методические рекомендации по выполнению практических работ по дисциплине «Автоматизированная обработка и перевод текста» разработаны в соответствии с рабочей программой дисциплины по направлению 45.04.02 Лингвистика, профиль – Современные методы прикладной лингвистики и перевода.

Практические задания разработаны в соответствии с рабочей программой дисциплины «Автоматизированная обработка и перевод текста», целью которой является развитие практических навыков в области использования автоматизированных систем анализа текста на естественных языках посредством современных цифровых компьютерных технологий, а также развитие профессиональных навыков в области эксплуатации программных сред автоматизированного перевода текста на естественных языках, призванных обеспечить решение типовых задач прикладной и цифровой лингвистики.

Целью практических занятий является закрепление теоретических знаний и приобретение практических умений и навыков, необходимых для освоения базовых принципов автоматизированного анализа и перевода текстов.

Методические рекомендации по каждой практической работе имеют теоретическую часть, необходимую для выполнения практических заданий. Практические задания органично сочетаются с теоретическими знаниями.

## Практическое занятие 1-2.

**Тема:** Общие принципы применения автоматизированных систем анализа текста.

**Цель:** Систематизация знаний в области применения автоматизированных систем анализа текста.

### Реализуемые компетенции:

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

## ВОПРОСЫ И ЗАДАНИЯ

1) Изучите следующий формат описания синтагматических конструкций.

№ №	Условные обозначения	Значения
1	2	3
1	*	Соответствует любому количеству произвольных лексических единиц и знаков препинания кроме терминальных (завершающих предложение)
2	\\	Соответствует границе предложения. Если предложение содержит прямую речь, то прямая речь выделяется как отдельное предложение или совокупность предложений
3	[*]	Соответствует одной произвольной лексической единице
4	[элемент]	Соответствует приведенному в скобках слову, нетерминальному знаку препинания или словосочетанию. Указывается буквами в нижнем регистре, например: [like]
5	[ГРУППА]	Соответствует группе слов, приведенной в квадратных скобках. Указывается буквами в верхнем регистре, например: [PERSONAL_PRONOUN] – соответствует одному любому личному местоимению
6	<условие>	Условие, указанное в угловых скобках, является необязательным для выполнения, например: <> - соответствует наличию или отсутствию запятой
7	{эл.1 эл.2 ... эл.N}	Соответствует одному любому из приведенных в фигурных скобках равновероятных элементов, например: {you know y'know}
8	(знак)	Соответствует указанному в скобках терминальному знаку препинания, например: (.)
9	(зн.1 зн.2 ... зн.N)	Соответствует одному из указанных терминальных знаков препинания, например: (. ... !)

Под терминальными понимаются знаки препинания, используемые для завершения предложения (точка, восклицательный знак, вопросительный знак, многоточие, а также их комбинации). К нетерминальным, напротив, относятся знаки препинания, которые не могут быть использованы для завершения предложения (запятая, двоеточие, дефис и т.п.).

2) Распишите следующие формализованные представления регулярных выражений с помощью метасимволов Perl.

\\ [why] [,] \* (.) \\

Данное регулярное выражение трактуется следующим образом: в начале предложения должно находиться слово *why*, отделенное запятой. За ним может следовать любое количество произвольных лексических единиц и нетерминальных знаков препинания, после чего предложение должно заканчиваться точкой.

Важным представляется то, что слова и словосочетания в квадратных скобках всегда указываются со строчной буквы. Это связано с тем, что автоматизированный лексический анализатор текста перед сравнением предложения с шаблоном приводит все слова к нижнему регистру, чтобы обеспечить независимость хода анализа от регистра, в котором написаны слова в предложении. В случае, если регистр, в котором написано то или иное слово, может повлиять на результат анализа, на этапе предварительной обработки предложения проводится дополнительный, чувствительный к регистру символов анализ.

Итак, регулярному выражению 1) будут соответствовать, например, следующие предложения:

а) *Why*, I think it's quite obvious. (Из разг. речи).

б) *Why*, it's very interesting. (Из разг. речи).

Предложение «*Why*, of course!», напротив, не соответствует регулярному выражению 1), так как терминальный знак препинания в данном случае – восклицательный знак, а не точка.

В случае, если на определенном месте в предложении может находиться одна из нескольких равновероятных лексических единиц, что не влияет на результат анализа, используется условное выражение вида «{слово|слово|...}», например:

\\ [I] [,] {uh|um} \* (...) \\

Регулярному выражению 2) будут соответствовать следующие предложения:

а) *I, uh* ... don't even know where to start. (Из разг. речи).

б) *I, um* ... I'm not interested in computer animation. (Из разг. речи).

## Список рекомендуемой литературы

### Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментальный переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

### Дополнительная литература:

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

### Методическая литература

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

### Интернет-ресурсы

1. Онлайн-выставка Корпорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>
2. Applied Linguistics – <https://www.appliedlinguistics.org>
3. Center for Applied Linguistics – <https://www.cal.org>
4. LibreOffice – <https://www.libreoffice.org>
5. OmegaT – <https://www.omegat.org>
6. PortableApps – <https://www.portableapps.com>
7. Praat – <https://www.fon.hum.uva.nl/praat>

### Практическое занятие 3-4.

**Тема:** Регулярные выражения как теоретико-методологическое основание построения систем автоматизированного анализа текста.

**Цель:** Систематизация знаний о регулярных выражениях как теоретико-методологическом основании построения систем автоматизированного анализа текста.

#### Реализуемые компетенции:

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

### ВОПРОСЫ И ЗАДАНИЯ

Распишите следующие формализованные представления регулярных выражений с помощью метасимволов Perl.

#### 1. Дискурсный маркер *well*.

Дискурсный маркер *well* может быть в большинстве случаев идентифицирован по занимаемой им начальной позиции в предложении. Как правило, данный дискурсный маркер отделяется запятой, поскольку входит в отдельную интонационную группу.

В случае, если дискурсный маркер *well* находится в повествовательном предложении, то он может использоваться либо с целью смены темы разговора, либо для выражения собственной точки зрения, либо для смягчения высказывания. В данном случае на основную функцию дискурсного маркера может также наслаиваться фокусная функция. В общем виде регулярное выражение, описывающее дискурсный маркер *well* в приведенных функциях, может быть представлено в следующем виде:

1. `\\ [well] [,] * (.) \\`

Например:

а) *Well*, that's no big deal for me, I don't really care. (Из разг. речи).

б) *Well*, let's speak about your recent promotion! (Из разг. речи).

в) *Well*, as I was saying before, it's not too important. (Из разг. речи).

Стоит отметить, что при выражении собственной точки зрения дискурсный маркер *well* может сочетаться с дискурсным маркером *why*, с которым он часто входит в одну интонационную группу, при этом дискурсный маркер *why* может, но не обязательно, быть отделен запятой. В данном случае для распознавания дискурсного маркера *well* применяется совокупность регулярных выражений вида:

2. `\\ [well] [why] [,] * \\`

3. `\\ [well] [,] [why] [,] * \\`

Например:

а) A: Do you think we could just go inside and grab the book?

B: *Well why*, of course, it's going to be really easy! (Из разг. речи).

б) A: I don't think it's an easy task.

B: *Well, why*, on the contrary – it's a piece of cake! (Из разг. речи).

В случае, если дискурсный маркер *well* используется с целью смягчения отрицательного ответа на вопрос, то до предложения с дискурсным маркером, как правило, следует вопросительное предложение, что может быть выражено с помощью регулярного выражения следующего вида:

4. \\ \* (?) \\ [well] [,] \* (.) \\

Например:

а) A: Can I borrow your textbook?

B: *Well*, I need it tonight in order to study. (Из разг. речи).

б) A: Do you have any classes tomorrow?

B: *Well*, yes, but nothing too important. (Из разг. речи).

В случае, если *well* используется для выражения недовольства или побуждения к действию, то он употребляется в восклицательном предложении, и определяется по регулярному выражению вида:

5. \\ [well] [,] \* (!) \\

Например:

а) *Well*, I don't think you're supposed to be here, go away! (Из разг. речи).

б) *Well*, I told you that you had to be careful! (Из разг. Речи).

## Список рекомендуемой литературы

### Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

### Дополнительная литература:

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

### Методическая литература

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

### Интернет-ресурсы

1. Онлайн-выставка Корсорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>

2. Applied Linguistics – <https://www.appliedlinguistics.org>

3. Center for Applied Linguistics – <https://www.cal.org>

4. LibreOffice – <https://www.libreoffice.org>
5. OmegaT – <https://www.omegat.org>
6. PortableApps – <https://www.portableapps.com>
7. Praat – <https://www.fon.hum.uva.nl/praat>

### Практическое занятие 5-6.

**Тема:** Нечеткие регулярные выражения. Расстояние Дамерау-Левенштейна.

**Цель:** Систематизация знаний в области нечетких регулярных выражений и нечеткого поиска текста.

#### Реализуемые компетенции:

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

### ВОПРОСЫ И ЗАДАНИЯ

Изучите нижеприведенные языковые модели и распишите их с применением метасимволов Perl.

Если дискурсный маркер *well* применяется как маркер планирования с целью обдумывания высказывания или ответа на вопрос, то за дискурсным маркером следует многоточие, при этом между дискурсным маркером и многоточием могут находиться несколько дополнительных лексических единиц. Стоит отметить, что при употреблении в данной функции дискурсный маркер может находиться в любой части предложения. Таким образом, для идентификации дискурсного маркера *well* в данной функции целесообразно использование следующей совокупности регулярных выражений, первое из которых является более точным, чем второе:

6. `\\ * [well] (...)`

7. `\\ * [well] * (...)`

Например:

а) *Well...* I don't know, I'm not sure... (Из разг. речи).

б) *Well I...* I don't know, I'm not sure... (Из разг. речи).

в) I want to say... *well...*, *well I...* No, I can't say this! (Из разг. речи).

Использование двух регулярных выражений диктуется необходимостью достижения максимально возможной точности определения дискурсных маркеров в тексте. Таким образом, поиск по регулярному выражению 7) осуществляется только в случае, когда предложение не соответствует регулярному выражению 6), то есть предложение не совпадает с более точным регулярным выражением.

В случае, если слово *well* находится в начальной позиции в предложении, но предложение не соответствует регулярным выражениям 1) – 7) и, таким образом, однозначно идентифицировать дискурсивный статус слова *well* не представляется возможным, подобное употребление слова *well* может быть отслежено по желанию пользователя с целью дальнейшего исследования. Таким образом, после проверок на соответствие предложения регулярным выражениям 1) – 7) выполняется проверка на соответствие регулярному выражению вида:

8. `\\ [well] * \\`

### Список рекомендуемой литературы

#### Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

#### **Дополнительная литература:**

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

#### **Методическая литература**

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

#### **Интернет-ресурсы**

1. Онлайн-выставка Корпорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>

2. Applied Linguistics – <https://www.appliedlinguistics.org>

3. Center for Applied Linguistics – <https://www.cal.org>

4. LibreOffice – <https://www.libreoffice.org>

5. OmegaT – <https://www.omegat.org>

6. PortableApps – <https://www.portableapps.com>

7. Praat – <https://www.fon.hum.uva.nl/praat>

### **Практическое занятие 7-8.**

**Тема:** Автоматизация обработки текста в редакторе vi/vim.

**Цель:** Систематизация знаний в области автоматизированной обработки текста в редакторе vi/vim.

#### **Реализуемые компетенции:**

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

### **ВОПРОСЫ И ЗАДАНИЯ**

Изучите нижеприведенные языковые модели и распишите их с применением метасимволов Perl.

2. *Дискурсный маркер oh.*

Слово *oh* независимо от позиции в предложении является дискурсным маркером, что

облегчает задачу его автоматизированного распознавания. В общем случае дискурсный маркер *oh* распознается по регулярному выражению вида:

- $\backslash \ * \ [oh] \ * \ \backslash$

В случае, если дискурсный маркер *oh* используется с целью указания на неточность приводимой информации или на ее приблизительный характер, или же употребляется с целью коррекции предыдущего высказывания, маркер может находиться между уточняемой и уточненной частями предложения и либо выделяется запятой с двух сторон, либо находится за многоточием и выделяется запятой спереди. Предложение с дискурсным маркером *oh* в указанной функции может быть повествовательным или восклицательным, при этом семантическая завершенность или незавершенность предложения не влияет на функцию дискурсного маркера. Регулярные выражения, описывающие приведенные ситуации, могут быть представлены в виде:

- $\backslash \ [*] \ * \ [,] \ [oh] \ [,] \ * \ (.\ \dots\ !)\ \backslash$
- $\backslash \ * \ (\dots)\ \backslash \ [*] \ [oh] \ [,] \ * \ (.\ \dots\ !)\ \backslash$

Например:

а) I think it was four days ago, *oh*, actually... maybe it was three days ago. (Из разг. речи).

б) I think it was in 1972... *oh*, no, probably in 1973. (Из разг. речи).

Если дискурсный маркер *oh* находится в начале предложения, он также отделяется запятой. Функция маркера *oh* зависит от типа предложения. Так, если предложение является повествовательным или восклицательным, то маркер может быть использован либо для введения уточненной или корректированной информации, либо для обеспечения обратной связи. Данная ситуация описывается с помощью регулярного выражения вида:

- $\backslash \ [oh] \ [,] \ * \ (.\ \dots\ !)\ \backslash$

Например:

а) I'm quite sure it was Neal who passed the test with the best score. *Oh*, wait... no, it was actually Alex. (Из разг. речи).

б) A: Actually, Dorian was present at Jack's birthday party – Jack finally decided to invite him!

B: *Oh*, that's interesting news! (Из разг. речи).

В противном случае, если предложение является вопросительным, то дискурсный маркер *oh* используется либо для запроса дополнительной информации коммуникантом, либо для фокусирования внимания реципиента на предложении, либо сочетает обе эти функции:

- $\backslash \ [oh] \ [,] \ * \ (?) \ \backslash$

Например:

а) *Oh*, do you know what this book is about? (Из разг. речи).

б) *Oh*, did you know that Ellen has written her new poem today? (Из разг. речи).

## Список рекомендуемой литературы

### Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

#### **Дополнительная литература:**

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

#### **Методическая литература**

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

#### **Интернет-ресурсы**

1. Онлайн-выставка Корсориума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>

2. Applied Linguistics – <https://www.appliedlinguistics.org>

3. Center for Applied Linguistics – <https://www.cal.org>

4. LibreOffice – <https://www.libreoffice.org>

5. OmegaT – <https://www.omegat.org>

6. PortableApps – <https://www.portableapps.com>

7. Praat – <https://www.fon.hum.uva.nl/praat>

### **Практическое занятие 9-10.**

**Тема:** Среда автоматизированной обработки текста GATE.

**Цель:** Систематизация знаний об автоматизированной обработке электронных корпусов текстов в среде GATE.

#### **Реализуемые компетенции:**

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

General Architecture for Text Engineering (GATE, программа) — система обработки естественного языка с открытым исходным кодом, использующая наборы компонентов на языке Java. Система изначально была разработана в Университете Шеффилда и развивается с 1995 г.

С помощью GATE реализуются задачи, где требуется выявить смысловое содержание текста и кодировать его в структурированном виде путём добавления аннотаций к сегментам текста. GATE используется наряду с NLTK, R и RapidMiner. Система применяется для извлечения информации, ручной и автоматической семантической аннотации, анализа кореферентности, работы с онтологиями (например, WordNet), машинного обучения (Weka, RASP, MAXENT, SVM Light), анализа потока сообщений в блогах (например, Twitter).

Семейство инструментов GATE включает: GATE Developer, GATE Mimir, GATE Cloud (для работы с крупномасштабными лингвистическими проектами), GATE Teamware (оптимизация работы серверов для совместного аннотирования текстов), GATE Embedded (библиотека объектов).

GATE поддерживается обширным сообществом разработчиков, пользователей, преподавателей, студентов и учёных. Используется в коммерческих и научно-исследовательских проектах крупными корпорациями, научно-исследовательскими лабораториями и университетами, коммерческими предприятиями малого и среднего бизнеса по всему миру. GATE применяется в самых разных областях научных знаний, относящихся к Компьютерной лингвистике, Обработке естественного языка, Моделированию языковых процессов, вычислительной биологии и медицине. Проекты, использующие GATE: ForgetIT (UK), The National Archives (UK), EMILLE (UK), myGRID Архивная копия от 29 сентября 2013 на Wayback Machine (UK), АКТ (UK), KIT Semantic Platform, Ontotext (Bulgaria), MeManage (недоступная ссылка) (Germany), Med Dictate (Canada), IE Denso (Japan).

## **Архитектура и базовые операции**

Архитектура GATE состоит из взаимосвязанных компонентов: «кусочков» программного обеспечения с четко определенными интерфейсами, которые могут быть развернуты в различных контекстах. В GATE реализованы готовые решения для токенизации, тегирования, разделения текста на высказывания (сплитер), извлечения именованных сущностей, машинного обучения. Компоненты делятся на три категории по функциям:

Language Resources (LR) — лингвистические ресурсы (данные),  
Processing Resources (PR) — программы для обработки документов (ресурсы),  
Visual Resources (VR) — графические интерфейсы для LR и PR.

Поддерживаются следующие форматы документов: Plain Text, HTML, SGML, XML, RTF, Email, PDF (some documents), Microsoft Office (some formats), OpenOffice (some formats), UIMA CAS, CoNLL/IOB. Работа с форматами документов в GATE имеет ряд специфических особенностей. В GATE встроены различные средства для работы с Unicode. Поддерживает языки: английский (по умолчанию), испанский, китайский, арабский, болгарский, французский, немецкий, хинди, итальянский, кебуано, румынский, русский.

При запуске программы в её основном окне содержится четыре основных пункта меню: Applications, Language Resources, Processing Resources, Datastores.

Applications

Контроллер вместе с входящими в него программами для обработки текста (Processing Resources). Определенные и сохраненные процессы обработки текста можно применить повторно к отдельному документу или корпусу текстов. Это обеспечивает надежную обработку текста и экономит время.

### **Language Resources (LR)**

Содержит три типа данных: документы, корпуса и графы аннотаций.

Document/ Blank Document — документ Gate, загруженный из файла или пустой. Новый документ создается через Language Resources > New > Gate Document. Документ можно сохранить в формате XML (правой кнопкой по названию документа > Save to XML).

Annotation Graph

Gate Corpus — корпус для хранения документов. Корпус создаётся через Language Resources > New > Gate Corpus. Наполнить корпус можно, указав список документов при создании или добавив документы в интерфейсе уже созданного корпуса, или с помощью команды Populate. Корпус можно сохранить в XML тем же способом, но надо указать не имя файла, а папку.

Аннотации организованы в виде графов, которые моделируются как Java-наборы. Аннотации представлены в виде дуг с начальным и конечным узлами, ID, присвоенным типом и FeatureMap (набором объектов). Узлы содержат указатели на источники в документе.

## **Processing Resources (PR)**

Программы для обработки текстов. В GATE ресурсы используются для автоматического создания и управления аннотациями. С помощью PR можно добавить или изменить разметку документа. Создается новый PR тем же способом, что и LR. При создании PR задаются параметры, которые бывают двух типов: параметры инициализации и параметры запуска. Первые требуется задать при создании ресурса, вторые — непосредственно перед его запуском из контроллера. Контроллеры управляют работой PR. Они отвечают за порядок применения PR и взаимодействие PR с LR. Основные типы контроллеров:

Pipeline. Последовательное применение цепочки PR к документу. После инициализации нужных ресурсов и добавления документов, можно создать контроллер и запустить наши ресурсы в заданной последовательности с заданными параметрами на заданном наборе документов.

Corpus Pipeline. Применение цепочки PR к корпусу.

Принципы работы с Pipeline и Corpus Pipeline похожи: создается новый контроллер (правой кнопкой на Applications > New > название контроллера), PR выбираются из списка слева и устанавливаются в заданном пользователем порядке. Указываются в случае Pipeline целевой документ, в случае Corpus Pipeline — целевой корпус, определяются параметры для PR. После запуска Run контроллер начнёт последовательно запускать PR на выбранных документах в порядке, заданном пользователем.

Конфигурации контроллеров (PR + настройки) можно сохранять в Applications (Save application state), желательно используя расширение .gapp.

## **Набор ресурсов CREOLE GATE CREOLE Plugin-manager**

Набор ресурсов, интегрированных с GATE, известен как CREOLE — Reusable Objects for Language Engineering. Ресурсы хранятся в CREOLE репозиториях, которые содержат XML файлы, Java-архив коды и необходимые для ресурсов библиотеки. Используемые ресурсы группируются вместе в плагины, которые хранятся по определенному адресу (URL или file:URL). Плагины бывают базовыми (загружаются при инсталляции GATE) и пользовательскими, могут находиться на локальном диске или удаленном сервере. Когда файл загружается в GATE, он выглядит как файл настроек creole.xml. Плагины CREOLE управляются через интерфейс с помощью команды File > Manage CREOLE Plugins > Add new creole repository. GATE представляется как набор настроек (функций), куда пользователь подключает компоненты CREOLE: пользователь задает список адресов, GATE извлекает по ним соответствующие ресурсы (PR). При выборе плагина в поле справа появляется его список настроек.

## Datastores

Хранилище данных. Необходим для хранения документов/корпусов и процессов для дальнейшего использования. Из всех видов хранилищ часто используется Serial DataStore. Хранилище можно

- создавать (Datastores > Create datastore > Serial DataStore > указать пустую папку без кириллицы и пробелов в пути),
- открывать ранее созданное (Open datastore),
- сохранять в него документы и корпуса (Open datastore > двойной щелчок на документе/корпусе),
- сохранять сделанные в документе или корпусе изменения (Save to its datastore).

Последовательность процессов обработки текстов из Applications можно запустить из-под DataStore. Нужно открыть DataStore, открыть корпус, а затем в Applications выбрать этот корпус. При запуске Application на корпусе из DataStore каждый документ будет загружен, обработан, сохранен и закрыт. То есть, в одно и то же время обрабатывается только один документ. Это не перегружает память, но процесс идет медленнее, чем если бы обрабатывались все документы одновременно.

## Разметка документов в GATE Developer

Разметка документов в соответствии с едиными правилами позволяет реализовать поиск и извлечение данных из документа, создавать онтологии.

В GATE после открытия редактора документов появляются вкладки Annotation Sets и Annotations List (или Annotations в зависимости от версии Gate), где в списке справа можно отметить галочкой типы аннотаций, которые требуется отобразить, или создать аннотации. Предусмотрена функция Change Color. Когда выделена часть текста, которой нужно присвоить аннотацию, появляется окно редактора аннотаций, где содержится следующие поля и элементы управления:

Annotations Editor Window

Тип аннотации (если какие-то типы уже были добавлены, можно выбрать из существующих)

Название атрибута (пустые поля для следующего атрибута появляются автоматически)

Значение атрибута

Кнопка удаления аннотации

Аннотации группируются в AnnotationSet-ы. Это удобная функция, которая позволяет хранить для одного документа несколько вариантов разметки, например, экспертную и автоматическую. Экспертная разметка, как правило, хранится в AnnotationSet под названием Key. Автоматическая разметка, как правило, пишется в пустой AnnotationSet, который присутствует во всех документах по умолчанию. В документах не присутствует AnnotationSet Key, его нужно создать, ввести в поле под типами аннотаций слово Key и нажать New.

## Система ANNIE

GATE располагает системой ИИ под названием ANNIE (A Nearly-New Information Extraction System), которая включает в себя набор ресурсов, обеспечивающих токенизацию (ANNIE English Tokenizer), POS-тэггинг (ANNIE POS-Tagger), разбивку на предложения

(ANNIE Sentence Splitter), извлечение именованных сущностей (ANNIE Gazetteer и ANNIE NE Transducer) и анализ кореферентности (ANNIE OrthoMatcher). Разработчики: Hamish Cunningham, Valentin Tablan, Diana Maynard, Kalina Bontcheva, Marin Dimitrov и другие. Разработчики ANNIE используют алгоритмы конечных автоматов и регулярные выражения JAPE.

Компоненты ANNIE объединены в приложение, поэтому для инициализации достаточно нажать на соответствующую зелёную иконку на панели инструментов GATE и выбрать with defaults. После того, как все ресурсы появятся в списке Processing Resources, двойным щелчком открывается приложение ANNIE, которое находится в списке Applications. Интерфейс приложения ANNIE такой же, как и у остальных приложений. В правой части — упорядоченный список ресурсов, которые будут вызываться на документе именно в той последовательности, которая задана в списке. Приложение ANNIE относится к классу Corpus Pipeline, то есть оно должно запускаться на корпусе текстов. В результате работы системы ANNIE в дефолтный AnnotationSet добавляется ряд аннотаций, в том числе Token (токены), Sentence (предложения), Lookup (словарные входы), Person, Location, Organization. Если корпус и документы перед запуском ANNIE лежали в DataStore, ANNIE будет извлекать документы по одному, обрабатывать их и складывать обратно.

### **Список компонентов:**

В списке PR есть ресурс Document Reset PR, который удаляет разметку из документа перед обработкой. У ресурса есть параметр setsToKeep, в котором перечисляются названия AnnotationSet-ов, которые не нужно очищать. При наличии в документе ручной разметки необходимо убедиться, что AnnotationSet, содержащий эту разметку, указан в этом списке, иначе он будет стёрт. По умолчанию там указан именно set Key.

Tokenizer. Делит текст на токены, а именно числа, знаки пунктуации, слова, символы, знаки пробелов. Каждому токenu присваивается его тип, соответственно, Word (атрибут orth с параметрами: upperInitial, allCaps, lowerCase, mixedCaps), Number, Symbol, Punctuation, SpaceToken.

Gazetteer. Определяет именованные сущности в соответствии со списком, то есть текстовым файлом (plain text) с перечислением именованных сущностей построчно. Каждый список содержит различные множества именованных сущностей: города, организации, дни недели и т. д. только для одного из заданных языков. Файл lists.def открывает доступ к списку именованных сущностей. Можно задать тип аннотации для отдельного списка именованных сущностей. Каждый список должен находиться в том же каталоге, что и индекс-файл lists.def.

Sentence Splitter. Делит текст на высказывания. Splitter использует список аббревиатур из Gazetteer, чтобы отличить конец высказывания от других типов использования знаков препинания. Каждому высказыванию присваивается аннотация 'Sentence', внутри которой находится аннотация 'Split' со значением 'internal' и 'external' для различия вопросительных, восклицательных, утвердительных предложений.

RegEx Sentence Splitter. Альтернативный способ деления текста на высказывания с помощью регулярных выражений JAPE.

Part of Speech Tagger. Проводит аннотацию каждого слова и символа. Существует список используемых тегов. Использует лексику и множество правил, сформированных на корпусе из Wall Street Journal. Правила и лексику можно менять вручную.

Semantic Tagger. Семантическая аннотация осуществляется с помощью правил JAPE-преобразователя (регулярных выражений), которые используют теги, полученные на предыдущих этапах аннотирования.

Orthographic Coreference (OrthoMatcher или 'NameMatcher'). Добавляет тип отношений между тегами именованных сущностей и семантическими тегами. Не находит новые

именованные сущности, но может приписать неклассифицируемый тип имени собственного, ориентируясь на совпадения слов.

Pronominal Coreference. Подключается в PR как дополнительный ресурс плагина ANNIE. Может находить цитируемый текст, прямую речь, местоимения (замена существительных, анафора), повторы. Требуется предварительное аннотирование по тегам: Token (English Tokenizer), Sentence (Sentence Splitter), Split (Sentence Splitter), Location (NE Transducer, OrthoMatcher), Person (NE Transducer, OrthoMatcher), Organization (NE Transducer, OrthoMatcher). Анализ осуществляется с помощью правил JARE-преобразователя (регулярных выражений), строится конечный автомат для поиска знаков прямой речи и цитат (одинарных и двойных кавычек, др.), поиска местоимений «it, its, itself» и «I, me, my, myself».

### **Тестирование результатов обработки текста**

В GATE встроены готовые инструменты для тестирования результатов обработки текста:

AnnotationDiff Tool выполняет сравнение на одном документе (кнопка сравнения на панели GATE).

Параметры: Key set (имя AnnotationSet с разметкой № 1), Response set (имя AnnotationSet с разметкой № 2), Type (одновременно можно тестировать только один тип), Features (атрибуты, которые мы сравниваем). После запуска сравнения (Compare) AnnotationDiff выведет пары аннотаций, отметив цветом тип пары. Типы: Correct (полное совпадение), Missing (правильная аннотация не найдена), Spurious/False positive (найдена лишняя аннотация), Partially correct (частичное пересечение границ аннотаций). В соответствии с количеством пар разных типов вычисляются стандартные метрики Precision, Recall и F.

Corpus Quality Assurance (CQA) предназначен для вычисления метрик качества на корпусе.

В окне нужного корпуса открывается вкладка Corpus Quality Assurance. CQA соберет данные по сетам, аннотациям и их атрибутам. Необходимо задать параметры, выбрать типы оценок (например, F1.0-score strict, lenient и average). Запустить сравнение. На вкладке Document Statistics можно посмотреть статистику по документам, а также сразу открыть выбранный документ или посмотреть для него AnnotationDiff. Результаты тестирования можно экспортировать в HTML.

### **GATE Mimir**

GATE Mimir — система обработки поисковых запросов с открытым исходным кодом, размещенным на SourceForge, распространяемая по GNU Lesser General Public Licence 3.0. GATE Mimir обеспечивает поддержку индексации и поиск языковой и семантической информации, поступающей из приложений с огромными базами лингвистических данных. GATE Mimir дает возможность поиска информации по тексту, аннотациям, семантическим онтологиям и семантическим метаданным, используя произвольные комбинации текста, структурную информацию и SPARQL. Разработчики не предоставляют готовый укомплектованный продукт, но предлагают скачать исходники и скомпилировать ваши собственные файлы. Предоставляется инструкция пользователя GATE Mimir, примеры использования, для последних версий релиза доступны архивы целого дерева исходников.

## **Список рекомендуемой литературы**

### **Основная литература**

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

#### **Дополнительная литература:**

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

#### **Методическая литература**

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

#### **Интернет-ресурсы**

1. Онлайн-выставка Корсорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>

2. Applied Linguistics – <https://www.appliedlinguistics.org>

3. Center for Applied Linguistics – <https://www.cal.org>

4. LibreOffice – <https://www.libreoffice.org>

5. OmegaT – <https://www.omegat.org>

6. PortableApps – <https://www.portableapps.com>

7. Praat – <https://www.fon.hum.uva.nl/praat>

### **Практическое занятие 11-12.**

**Тема:** Среда автоматизированного перевода текста OmegaT.

**Цель:** Систематизация знаний в области применения среды автоматизированного перевода текста OmegaT.

#### **Реализуемые компетенции:**

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

OmegaT — это свободная программа перевода, поддерживающая память переводов и работающая на Windows, macOS, Linux... Эта система предназначена для профессиональных переводчиков. Она не переводит вместо человека! (В отличие от программ, выполняющих «машинный перевод», OmegaT лишь помогает переводчику и упрощает его работу.)

Профессиональные функции  
подбор неточных совпадений;  
размножение совпадений;  
одновременная обработка проектов с большим числом файлов;  
одновременное использование нескольких памятей переводов;  
использование глоссариев с распознаванием флективных форм;

Поддержка более 30 форматов файлов  
Microsoft Word, Excel, Power;  
XHTML и HTML;  
Open Document Format (LibreOffice, OpenOffice.org);  
MediaWiki (Wikipedia);  
неформатированный текст;  
около 30 других форматов файлов;

Эффективные инструменты  
поддержка юникода (UTF-8): используется для нелатинских алфавитов;  
поддержка языков с письмом справа налево;  
встроенное средство проверки орфографии;  
совместимость с другими программами автоматизированного перевода (TMX, TTX, TXML, XLIFF, SDLXLIFF);  
интерфейс для Google Translate

## **Список рекомендуемой литературы**

### **Основная литература**

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

### **Дополнительная литература:**

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

### **Методическая литература**

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

### **Интернет-ресурсы**

1. Онлайн-выставка Корсорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>
2. Applied Linguistics – <https://www.appliedlinguistics.org>
3. Center for Applied Linguistics – <https://www.cal.org>
4. LibreOffice – <https://www.libreoffice.org>
5. OmegaT – <https://www.omegat.org>
6. PortableApps – <https://www.portableapps.com>
7. Praat – <https://www.fon.hum.uva.nl/praat>

### **Практическое занятие 13-14.**

**Тема:** Инструментарий автоматизированного перевода. Память переводов, глоссарий, API машинного перевода.

**Цель:** Систематизация знаний в области использования профессионального инструментария систем автоматизированного перевода текстов (САТ).

**Реализуемые компетенции:**

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

**Актуальность:** актуальность изучения темы определяется практической целесообразностью использования лингвистических алгоритмов для оптимизации деятельности лингвиста-практика в части письменного перевода деловой документации.

**Память переводов (ТМ)** — это база данных ранее переведенного текста и ключевой компонент переводческих инструментов. Памяти переводов состоят из целых переведенных сегментов или фраз, в то время как термины — это отдельные слова.

Основные преимущества использования ТМ:

ТМ позволяет повторно использовать переводы. Это ускоряет процесс перевода и снижает затраты.

ТМ помогает обеспечить согласованность перевода. Это важно, когда у клиента есть более одного переводчика, работающего над проектом.

ТМ служит долгосрочной резервной копией, когда проект или задание больше не доступны.

Данные перевода могут обмениваться между Phrase TMS и Phrase Strings, экспортируя переводы в формате .TMX из любого приложения. Чтобы экспортировать из Strings, скачайте переводы для нужного языка из проекта в формате .TMX.

Импорт напрямую в ТМ не поддерживается в Strings. Вместо этого файл .TMX или .XLSX может быть импортирован как ключи в проект, который затем может делиться своим контентом с организационной памятью переводов.

Память переводов Strings основана на проекте и имеет меньше метаданных, чем память переводов TMS.

TMS

Тексты разделяются на более мелкие сегменты (обычно предложения или заголовки) во время сегментации. Оригинальный сегмент и его перевод затем сохраняются в памяти переводов как единица перевода, когда этот сегмент подтвержден пользователем в редакторе.

Сегменты, которые автоматически подтверждаются на основе предварительного перевода настроек, не сохраняются, так как они уже находятся в выбранной ТМ, но контекст добавляется в существующую ТМ.

Если ТМ создана и добавлена в существующий проект, подтвержденные сегменты из этого проекта будут добавлены в новую ТМ. Настоятельно рекомендуется, чтобы лингвисты подтверждали каждый сегмент, в который они вложили усилия, прежде чем переходить к другому сегменту, чтобы гарантировать, что усилия не потеряны.

ТМ может иметь несколько целевых языков, но только один язык источника и может использоваться в нескольких проектах одновременно.

Хотя использование памяти переводов настоятельно рекомендуется, существуют некоторые ограничения:

Хотя нет конкретных ограничений на количество целевых языков, которые может иметь память переводов, или на количество единиц перевода, которые могут быть сохранены в памяти переводов, очень большие памяти переводов (миллионы сегментов) замедляют производительность поиска, предварительного перевода или анализа и могут быть трудными для поддержки и редактирования.

Проект может иметь несколько памяти переводов, поэтому всегда лучше иметь несколько меньших, хорошо управляемых памяти переводов, чем одну очень большую.

Максимум 10 памяти переводов может быть назначено любому проекту/языковой паре на каждом этапе рабочего процесса.

Память переводов (ТМ) — это база данных ранее переведенного текста, используемая в проектах, и она предлагает совпадающие результаты при переводе контента, основываясь на языке по умолчанию проекта. Это ускоряет процесс перевода и улучшает согласованность текстов и проектов.

ТМ является инструментом перевода для всей организации, который может быть назначен различным проектам. В противном случае, если ни один из проектов не использует память переводов, каждый проект будет полагаться на свою собственную локальную память переводов.

Обновление памяти переводов не требуется, так как она автоматически обновляется с каждым новым переводом.

## **Список рекомендуемой литературы**

### **Основная литература**

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Баймуратова, У. Электронный инструментарий переводчика : учебное пособие / У. Баймуратова ; Министерство образования и науки Российской Федерации, Федеральное государ-

ственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. 5th GATE Training Course. - The University of Sheffield, 2012. - <http://gate.ac.uk/wiki/TrainingCourseJune2012/>

#### **Дополнительная литература:**

1. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.

2. Поршнева, Е.Р. Базовая лингвистическая подготовка переводчика: монография / Е.Р. Поршнева. — Нижний Новгород: ННГУ им. Н.И. Лобачевского, 2002. — 148 с.

3. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

#### **Методическая литература**

1. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Автоматизированная обработка и перевод текста». – Ставрополь, СКФУ, 2026. (электронная версия).

#### **Интернет-ресурсы**

1. Онлайн-выставка Корсорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” – <https://narfu.ru/virtualika/>

2. Applied Linguistics – <https://www.appliedlinguistics.org>

3. Center for Applied Linguistics – <https://www.cal.org>

4. LibreOffice – <https://www.libreoffice.org>

5. OmegaT – <https://www.omegat.org>

6. PortableApps – <https://www.portableapps.com>

7. Praat – <https://www.fon.hum.uva.nl/praat>

## **Методические указания**

по выполнению самостоятельной работы студентов по дисциплине  
«Автоматизированная обработка и перевод текста»

Направление подготовки	<u>45.04.02 Лингвистика</u>
Направленность (профиль)	<u>Современные методы прикладной лингвистики и перевода</u>
Год начала подготовки	<u>2026</u>
Форма обучения	<u>Очная</u>
Реализуется в семестре	<u>2</u>

**Ставрополь**  
**2026**

## Содержание

1. Введение .....	3
2. Общая характеристика самостоятельной работы студента при изучении дисциплины.....	4
3. План-график выполнения самостоятельной работы .....	5
4. Методические рекомендации по изучению теоретического материала .....	5
5. Методические указания (по видам работ, предусмотренных рабочей программой дисциплины).....	6
6. Список литературы, использованной при составлении методических рекомендаций .....	10

## 1. Введение

Методические рекомендации к самостоятельной работе магистрантов по дисциплине «Автоматизированная обработка и перевод текста» разработаны в соответствии с рабочей программой дисциплины по направлению 45.04.02 - Лингвистика, программа – Современные методы прикладной лингвистики и перевода.

Основной формой работы магистранта является не только работа на практических занятиях, их дополнение рекомендованной литературой, но и большая самостоятельная учебная работа, которая позволит глубоко проникнуть в суть рассматриваемой проблемы. Но для успешной учебной деятельности, ее интенсификации необходимо учитывать следующие субъективные факторы:

1. Знание программного материала, наличие прочной системы знаний, необходимой для усвоения основных дисциплин, предусмотренных программой, общая совокупность которых обуславливает уровень овладения грамматическим компонентом иноязычной речи.

2. Наличие выработанных умений, навыков умственного труда:

а) умение делать глубокий, обстоятельный анализ при работе с книгой, Интернет–источниками;

б) владение логическими операциями: сравнение, анализ, обобщение, определение понятий, правила систематизации и классификации.

3. Специфика познавательных психических процессов: внимание, память, речь, наблюдательность, интеллект и мышление.

4. Хорошая работоспособность, которая обеспечивается нормальным физическим состоянием.

5. Соответствие избранной деятельности, профессии индивидуальным способностям. Необходимо выработать умение саморегулировать свое эмоциональное состояние и устранять обстоятельства, нарушающие деловой настрой, мешающие намеченной работе.

6. Овладение оптимальным стилем работы, обеспечивающим успех в деятельности.

7. Уровень требований к себе, определяемый сложившейся самооценкой.

## 2. Общая характеристика самостоятельной работы студента при изучении дисциплины

Самостоятельная работа магистранта в рамках дисциплины «Основы программирования лингвистических алгоритмов» понимается как планируемая учебная работа, выполняемая во внеаудиторное (аудиторное) время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия.

Самостоятельная работа направлена на формирование следующих компетенций:

Индекс	Формулировка:
--------	---------------

УК-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
ОПК-7	Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

Цель самостоятельной работы магистрантов в процессе изучения дисциплины «Основы программирования лингвистических алгоритмов» – научить магистранта осмысленно и самостоятельно работать: 1) с учебным материалом по дисциплине, 2) с научной информацией, актуальными исследованиями в области лингвистики, 3) с эмпирическими данными, получаемыми в ходе экспериментальных лингвистических исследований, 4) с методологическими подходами современных лингвистических исследований; 5) с конкретными лингвистическими методами и методиками.

Задачи самостоятельной работы:

- систематизировать и закрепить полученные теоретические знания и практические умения магистрантов;

- развить познавательные способности и активность магистрантов: творческую инициативу, самостоятельность, ответственность и организованность;

- сформировать и развить навыки ведения самостоятельной работы и овладения методикой исследования при решении разрабатываемых в учебной деятельности проблем и вопросов;

- повысить уровень подготовленности к самостоятельной работе в соответствии с выбранным научным направлением в условиях современного состояния науки и культуры.

Таким образом, самостоятельная работа приобщает научному и исследовательскому творчеству, поиску и анализу актуальных проблем современной психолингвистической науки.

### 3. План-график выполнения самостоятельной работы.

Код реализуемой компетенции	Вид деятельности студентов	Итоговый продукт самостоятельной работы	Средства и технологии оценки	Объем часов
УК-6, ОПК-7	Изучение литературы	Конспект	Собеседование	20
УК-6, ОПК-7	Подготовка к круглому столу	Индивидуальное задание	Собеседование	20
УК-6, ОПК-7	Работа с электронными ресурсами в сети Интернет	Конспект	Собеседование	20
УК-6, ОПК-7	Подготовка доклада	Доклад и презентация	Собеседование	20
Итого за 2 семестр				80
Итого				80

Для выполнения самостоятельной работы необходимо пользоваться литературой, которая предложена в списке рекомендуемой литературы, Интернет-ресурсами или другими источниками по усмотрению магистранта.

Самостоятельная работа рассчитана на разные уровни мыслительной деятельности. Выполненная работа позволит приобрести не только знания, но и умения, навыки, а также выработать свою методику подготовки, что очень важно в дальнейшем процессе научной деятельности.

При изучении дисциплины предусматриваются следующие формы самостоятельной работы магистранта:

- самостоятельное изучение основной и дополнительной литературы по дисциплине с конспектированием по разделам;
- работа с электронными ресурсами в сети Интернет;
- конспектирование и реферирование первоисточника и научно-исследовательской литературы;
- подготовка к семинару-круглому столу;
- подготовка мультимедийной презентации;
- подготовка доклада.

### 4. Методические рекомендации по изучению теоретического материала

*Чтение основной и дополнительной литературы по курсу с конспектированием по разделам.*

Самостоятельная работа при чтении учебной литературы начинается с изучения конспекта материала, полученного при слушании лекций преподавателя. Полученную информацию необходимо осмыслить. При необходимости, в конспект лекций могут быть внесены схемы, другая дополнительная информация. При изучении нового материала составляется конспект. Сжато излагается самое существенное в данном материале.

*Работа с электронными ресурсами в сети Интернет.*

Для повышения эффективности самостоятельной работы магистрант должен уметь работать в поисковой системе сети Интернет и использовать найденную информацию при подготовке к занятиям. Поиск информации можно вести по автору, заглавию, виду издания, году издания или издательству. Также в сети Интернет доступна услуга по скачиванию методических указаний и учебных пособий, подбору необходимой научной литературы.

*Конспектирование и реферирование первоисточника и научно-исследовательской литературы.*

Конспект представляет собой дословные выписки из текста источника. При этом необходимо понимать, что конспект – это не полное переписывание чужого текста. Необходимо знать, что при написании конспекта сначала прочитывается текст – источник, в нём выделяются основные положения, подбираются примеры, идёт перекомпоновка материала, а уже затем оформляется текст конспекта. Конспект может быть полным, когда работа идёт со всем текстом источника или неполным, когда интерес представляет какой-либо один или несколько вопросов, затронутых в источнике.

Реферирование — это сложный творческий процесс, в основе которого лежит умение выделить главную информацию из текста первоисточника. Реферирование – процесс аналитически-синтетической обработки информации, которая заключается в анализе первичного документа, нахождении значимых в смысловом отношении данных (основных положений, фактов, доведите день, результатов, выводов) Реферирование имеет целью сократить физический объем первичного документа при сохранении его основного смыслового содержания, используется в научной, издательской, информационной и библиографической деятельности.

## **5 Методические указания (по видам работ, предусмотренных рабочей программой дисциплины)**

*Подготовка к круглому столу*

Подготовка к семинару-круглому столу начинается с распределение форм участия и функции магистрантов в семинаре-круглом столе. Магистрантами осуществляется определение круга проблем и вопросов, подлежащих обсуждению; подбор основной и дополнительной литературы к теме семинара - круглого стола, а также дальнейшее изучение литературы.

*Подготовка мультимедийной презентации*

Презентация, согласно толковому словарю русского языка Д.Н. Ушакова: «... способ подачи информации, в котором присутствуют рисунки, фотографии, анимация и звук». Для подготовки презентации рекомендуется использовать LibreOffice Impress (для подготовки собственно

мультимедийных презентаций) и LibreOffice Writer (для составления текстового сопровождения презентации), являющихся компонентами открытого и свободного офисного пакета LibreOffice. Также допускается использование проприетарного продукта Microsoft Office (Powerpoint и Word, соответственно), однако в этом случае должны использоваться наиболее совместимые форматы .pptx, .docx.

Для подготовки презентации необходимо собрать и обработать начальную информацию.

Последовательность подготовки презентации:

1. Четко сформулировать цель презентации: вы хотите свою аудиторию мотивировать, убедить, заразить какой-то идеей или просто формально отчитаться.
2. Определить каков будет формат презентации: живое выступление (тогда, сколько будет его продолжительность) или электронная рассылка (каков будет контекст презентации).
3. Отобрать всю содержательную часть для презентации и выстроить логическую цепочку представления.
4. Определить ключевые моменты в содержании текста и выделить их.
5. Определить виды визуализации (картинки) для отображения их на слайдах в соответствии с логикой, целью и спецификой материала.
6. Подобрать дизайн и форматировать слайды (количество картинок и текста, их расположение, цвет и размер).
7. Проверить визуальное восприятие презентации.

К видам визуализации относятся иллюстрации, образы, диаграммы, таблицы.

Иллюстрация – представление реально существующего зрительного ряда.

Образы – в отличие от иллюстраций – метафора. Их назначение – вызвать эмоцию и создать отношение к ней, воздействовать на аудиторию. С помощью хорошо продуманных и представляемых образов, информация может надолго остаться в памяти человека.

Диаграмма – визуализация количественных и качественных связей. Их используют для убедительной демонстрации данных, для пространственного мышления в дополнение к логическому.

Таблица – конкретный, наглядный и точный показ данных. Ее основное назначение – структурировать информацию, что порой облегчает восприятие данных аудиторией.

*Практические советы по подготовке презентации.*

- готовьте отдельно: печатный текст + слайды + раздаточный материал;
- слайды – визуальная подача информации, которая должна содержать
- минимум текста, максимум изображений, несущих смысловую нагрузку, выглядеть наглядно и просто;

- текстовое содержание презентации – устная речь или чтение, которая
- должна включать аргументы, факты, доказательства и эмоции;
- рекомендуемое число слайдов 10-12;
- обязательная информация для презентации: тема, фамилия и инициалы
- выступающего; план сообщения; краткие выводы из всего сказанного; список использованных источников;
- раздаточный материал – должен обеспечивать ту же глубину и охват, что и живое выступление: люди больше доверяют тому, что они могут унести с собой, чем исчезающим изображениям, слова и слайды забываются, а раздаточный материал остается постоянным осязаемым напоминанием; раздаточный материал важно раздавать в конце презентации; раздаточный материалы должны отличаться от слайдов, должны быть более информативными.

*Доклад*, согласно толковому словарю русского языка Д.Н. Ушакова:

«... сообщение по заданной теме, с целью внести знания из дополнительной литературы, систематизировать материал, проиллюстрировать примерами, развивать навыки самостоятельной работы с научной литературой, познавательный интерес к научному познанию».

Тема доклада должна быть согласована с преподавателем и соответствовать теме учебного занятия. Материалы при его подготовке, должны соответствовать научно-методическим требованиям вуза и быть указаны в докладе. Необходимо соблюдать регламент, оговоренный при получении задания. Иллюстрации должны быть достаточными, но не чрезмерными.

Работа магистранта над докладом-презентацией включает отработку умения самостоятельно обобщать материал и делать выводы в заключении, умения ориентироваться в материале и отвечать на дополнительные вопросы слушателей, отработку навыков ораторства, умения проводить диспут.

Докладчики должны знать и уметь: сообщать новую информацию; использовать технические средства; хорошо ориентироваться в теме всего семинарского занятия; дискутировать и быстро отвечать на заданные вопросы; четко выполнять установленный регламент (не более 10 минут); иметь представление о композиционной структуре доклада и др.

### *Структура выступления*

Вступление помогает обеспечить успех выступления по любой тематике. Вступление должно содержать: название, сообщение основной идеи, современную оценку предмета изложения, краткое перечисление рассматриваемых вопросов, живую интересную форму изложения, акцентирование внимания на важных моментах, оригинальность подхода.

Основная часть, в которой выступающий должен глубоко раскрыть суть

затронутой темы, обычно строится по принципу отчета. Задача основной части – представить достаточно данных для того, чтобы слушатели заинтересовались темой и захотели ознакомиться с материалами. При этом логическая структура теоретического блока не должны даваться без наглядных пособий, аудиовизуальных и визуальных материалов.

Заключение – ясное, четкое обобщение и краткие выводы, которых всегда ждут слушатели

### *Написание доклада*

Доклад – публичное сообщение, представляющее собой развёрнутое изложение определённой темы.

Этапы подготовки доклада:

1. Определение цели доклада.
2. Подбор необходимого материала, определяющего содержание доклада.
3. Составление плана доклада, распределение собранного материала в необходимой логической последовательности.
4. Общее знакомство с литературой и выделение среди источников главного.
5. Уточнение плана, отбор материала к каждому пункту плана.
6. Композиционное оформление доклада.
7. Заучивание, запоминание текста доклада, подготовки тезисов выступления.
8. Выступление с докладом.
9. Обсуждение доклада.
10. Оценивание доклада

Композиционное оформление доклада – это его реальная речевая внешняя структура, в ней отражается соотношение частей выступления по их цели, стилистическим особенностям, по объёму, сочетанию рациональных и эмоциональных моментов, как правило, элементами композиции доклада являются: вступление, определение предмета выступления, изложение (опровержение), заключение.

Вступление помогает обеспечить успех выступления по любой тематике.

Вступление должно содержать:

- название доклада;
- сообщение основной идеи;
- современную оценку предмета изложения;
- краткое перечисление рассматриваемых вопросов;
- интересную для слушателей форму изложения;
- акцентирование оригинальности подхода.

Выступление состоит из следующих частей:

Основная часть, в которой выступающий должен раскрыть суть темы, обычно строится по принципу отчёта. Задача основной части: представить

достаточно данных для того, чтобы слушатели заинтересовались темой и захотели ознакомиться с материалами.

Заключение - это чёткое обобщение и краткие выводы по излагаемой теме.

## **6. Список литературы, использованной при составлении методических рекомендаций**

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
2. Баймуратова, У. Электронный инструментарий переводчика: учебное пособие / У. Баймуратова; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2014. - 120 с. ; То же [Электронный ресурс]. - URL:<http://biblioclub.ru/index.php?page=book&id=259202>
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. — Ставрополь: Изд-во СКФУ, 2014. — 186 с.
4. Подготовка переводчиков: дисциплины по выбору: коллективная монография / под общ. ред. Н.Н. Гавриленко. — М.: Флинта, 2022. — 168 с.
5. Онлайн-выставка Корсорциума федеральных университетов “Цифровые практики в лингвистике и образовании” “Виртуалика” — <https://narfu.ru/virtualika/>
6. Applied Linguistics — <https://www.appliedlinguistics.org>
7. Center for Applied Linguistics — <https://www.cal.org>
8. LibreOffice — <https://www.libreoffice.org>
9. OmegaT — <https://www.omegat.org>
10. PortableApps — <https://www.portableapps.com>
11. Praat — <https://www.fon.hum.uva.nl/praat>