УДК 004.4

АНАЛИЗ ФРАНЦУЗСКОГО ЯЗЫКА С ПОМОЩЬЮ *PYTHON*

Ахрамович С.П., Сердюк К.В, Предченко В.М.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Лихачевский Д.В. – к. т. н., доцент, доцент кафедры ПИКС

Аннотация. В статье рассматривается анализ французского языка с использованием средств языка программирования *Python*. Проведен частотный и статистический анализ французских слов, выявлены самое длинное и короткое слова, а также рассчитана средняя длина слова. Кроме того, исследованы наиболее распространенные сочетания букв (биграммы) с применением библиотеки *pandas*.

Ключевые слова: частотный анализ, статический анализ, французский язык, длина слова.

Введение. Анализ естественного языка играет важную роль в лингвистических исследованиях и разработке технологий обработки текста. Исследование структуры слов и их характеристик позволяет лучше понимать закономерности языка, разрабатывать инструменты для машинного перевода и создания интеллектуальных систем. Французский язык, как один из наиболее распространенных мировых языков, представляет интерес с точки зрения анализа его статистических особенностей.

Основная часть. Для начала анализа французских слов необходимо было найти источник с французскими словами. Данные для проведения анализа были взяты из открытого репозитория на *GitHub*, содержащего порядка 336 тысяч слов французского языка [1].

Для обработки и анализа этих данных была использована библиотека *pandas* – быстрый, мощный, гибкий и простой в использовании инструмент с открытым исходным кодом для анализа и обработки больших объемов данных [2].

С помощью *pandas* данные были структурированы и подготовлены для дальнейшего статистического анализа. В ходе статического анализа слов французского языка был выполнен ряд задач, направленных на более глубокое понимание структуры и особенностей используемых слов и их сочетаний.

Одной из задач статического анализа было определение самого длинного слова французского языка. Для определения самого длинного слова использовалась функция *max* (рисунок 1).

```
longest_word = max(words, key=len)
print(f"Самое длинное слово: {longest_word}, длина: {len(longest_word)}")
```

Рисунок 1 – Определение самого длинного слова французского языка

Результатом выполнения данной функции является вывод самого длинного слова из данных – *anticonstitutionnellement*, которое состоит из 25 букв.

После определения самого длинного слова, следующей задачей в рамках статического анализа было определение средней длины слов. Для определения средней длины слова было использовано выражение (рисунок 2).

```
average_length = sum(len(word) for word in words) / len(words)
print(f"Средняя длина слова: {average_length:.2f}")
```

Рисунок 2 – Определение средней длины слова французского языка

Результатом выполнения данного кода является вывод среднего количества символов во французских словах – 10.09 символов.

Следующим этапом исследования стало определение самого короткого слова во французском языке. Для этого был использован метод нахождения минимального элемента в списке слов, основанный на их длине (рисунок 3).

```
shortest_word = min(words, key=len)
print(f"Самое короткое слово: {shortest_word}")
```

Рисунок 3 – Определение самого короткого слова французского языка

Результатом выполнения данного кода является вывод самого короткого слова – *a*.

После выполнения статического анализа, направленного на изучение структуры слов, следующим этапом является частотный анализ. В рамках частотного анализа были исследованы самые распространенные сочетания букв, что способствует более детальному изучению особенностей французского языка.

Ключевой задачей данного анализа стало определение наиболее распространенных сочетаний букв – биграмм (двухбуквенных сочетаний). Данный анализ позволяет выявить характерные языковые паттерны, закономерности в построении слов, часто встречающиеся во французских словах, а также определить распределение буквосочетаний, выявить отличительные черты орфографии. Биграммы рассчитывались путем перебора всех слов в файле, разбиения их на пары соседних букв и подсчета частоты встречаемости каждого сочетания (рисунок 4).

```
def get_bigrams(word_list):
    bigrams = Counter()
    for word in word_list:
        for i in range(len(word) - 1):
            bigram = word[i:i+2]
            bigrams[bigram] += 1
    return bigrams.most_common(10)

word_list = df["Word"].dropna().tolist()

bigram_freq = get_bigrams(word_list)
```

Рисунок 4 – Определение биграмм французского языка

Результатом выполнения данного кода является вывод биграмм (рисунок 5).

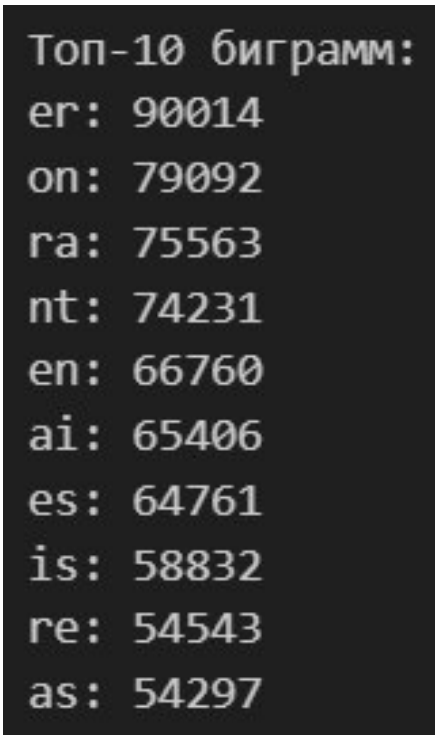

Рисунок 5 – Вывод наиболее часто используемых биграмм

Наиболее часто используемыми биграммами являются сочетания *er*, *on*, *ra*.

Заключение. В ходе проведенного исследования с использованием *Python* были выполнены статистический и частотный анализы французского языка. Статистический анализ показал, что самое длинное слово в французском языке – *anticonstitutionnellement* (25 символов), в то время как самое короткое – *a* (1 символ). Средняя длина слов составила 10,09 символов. Частотный анализ выявил наиболее распространенные биграммы (*er*, *on*, *ra*), что отражает характерные языковые паттерны и особенности французской орфографии. Полученные данные могут быть использованы для создания более точных инструментов обработки французского языка, таких как системы распознавания речи, анализаторы текста и интеллектуальные поисковые системы.

Список литературы

1 Words [Электронный ресурс]. – Режим доступа: https://github.com/lorenbrichter/Words/tree/master/Words. – Дата доступа: 03.03.2025.

2 Pandas [Электронный ресурс]. – Режим доступа: https://pandas.pydata.org/. – Дата доступа: 03.03.2025.

UDC 004.4

ANALYZING FRENCH LANGUAGE USING PYTHON

Akhramovich S.P., Siardziuk K.V., Predchenko V.M.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Likhachevsky D.V. – Cand. of Technical sciences, associate professor

Annotation. The article discusses methods for analyzing the French language using the Python programming language. A frequency and statistical analysis of French words was carried out, the longest and shortest words were identified, and the average word length was calculated. In addition, the most common letter combinations (bigrams) were studied using the pandas library.

Keywords: frequency analysis, static analysis, French, word length.