УДК 004.4

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЧАСТОТНОСТИ СЛОВ И СТРУКТУРНЫХ ОСОБЕННОСТЕЙ ФРАНЦУЗСКОГО И ИСПАНСКОГО ЯЗЫКОВ 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, декан ФКП

Аннотация. В статье представлен сравнительный анализ французского и испанского языков с использованием Python. Выполнен анализ лексем, их средней длины, а также исследуются биграммы с применением библиотеки pandas. На основе результатов анализа выявляются общие закономерности для обоих языков.

Ключевые слова: сравнительный анализ, лексема, французский язык, испанский язык, средняя длина

Введение. Сравнительный анализ языков, особенно таких распространенных, как французский и испанский, предоставляет уникальные возможности для выявления особенностей их структуры, лексического состава и частотного распределения слов. Применение вычислительных методов и алгоритмов обработки текста, в частности, с использованием языка программирования Python, значительно упрощает этот процесс и позволяет провести глубокий анализ.

Основная часть. Для начала проведения сравнительного анализа французских и испанских слов необходимо было найти источник со словами данных языков. Данные для проведения анализа были взяты из открытого репозитория на GitHub, содержащего около 336 тысяч слов французского языка и 636 тысяч слов испанского языка [1].

Для обработки и анализа данных была применена библиотека pandas – эффективный, универсальный и удобный инструмент с открытым исходным кодом, предназначенный для работы с большими объемами данных [2].

С использованием pandas данные были организованы и подготовлены для последующего статистического анализа. В процессе анализа слов французского и испанского языков была решена серия задач, направленных на более детальное исследование структуры слов и их сочетаний языков.

Одной из задач статистического анализа являлось определение самого длинного слова в французском и испанском языках. Для определения самого длинного слова использовалась функция max.

Проанализировав результаты, полученные при определении самого длинного слова в языках, можно отметить, что самое длинное слово в французском языке («anticonstitutionnellement», длина 25) значительно длиннее самого длинного слова в испанском языке («achicharronariamos», длина 18). Это может свидетельствовать о различиях в морфологической структуре языков: французский язык, с его множеством производных и сложных форм, склонен к образованию более длинных слов. В то время как в испанском языке, несмотря на наличие длинных слов, такие случаи встречаются реже.

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 1).

Результатом выполнения данного кода является вывод среднего количества символов во французских словах – 10.09 символов, в испанских лексемах составил 10.06. Сравнив результаты, стоит отметить, что средняя длина в двух языках практически идентична. Разница в 0.03 символа является минимальной, что указывает на схожие лексические характеристики

этих языков в контексте длины слов. Это может свидетельствовать о том, что, несмотря на различия в грамматической структуре и словообразовании, французский и испанский языки имеют схожие тенденции в использовании слов по длине.

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

Рисунок 1 – Определение средней длины слова французского и испанского языков

После завершения статического анализа данных был проведен частотный анализ. Данный анализ позволяет лучше понять языковые предпочтения и тенденции в употреблении слов, а также определить распределение комбинаций букв и выявить особенности орфографических конструкций.

Биграммы рассчитывались путем перебора всех слов в файле, разбиения их на пары соседних букв и подсчета частоты встречаемости каждого сочетания (рисунок 2).

```
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 2 – Определение биграмм французского языка

Самыми распространенными биграммами французского языка являются «*er*», «*on*», «*ra*» (рисунок 3).

```
Топ-10 биграмм:
er: 90014
on: 79092
ra: 75563
nt: 74231
en: 66760
ai: 65406
es: 64761
is: 58832
re: 54543
as: 54297
```

Рисунок 3 – Самые распространенные биграммы французского языка

Самыми распространенными биграммами испанского языка являются «*er*», «*on*», «*ra*» (рисунок 4).

```
Топ-10 биграмм:
ar: 220095
as: 170829
ra: 150604
es: 150431
re: 144188
en: 126933
os: 123867
is: 110088
an: 107323
ri: 100874
```

Рисунок 4 – Самые распространенные биграммы испанского языка

Оба языка имеют общие биграммы, такие как «ra», «es», «re», «en», «is» и «as», что указывает на схожие звуковые комбинации и, возможно, общие корни слов (оба языка романские). Частотность общих биграмм отличается. Например, «ar» является самой частотной

биграммой в испанском, но отсутствует в топ-10 французских биграмм. «Er» и «on», лидирующие во французском, отсутствуют в топ-10 испанских.

В испанском списке присутствуют «ar», «os», «an» и «ri», которых нет во французском. Во французском есть «er», «on», «nt» и «ai», отсутствующие в испанском. Это указывает на различия в фонологии и морфологии языков.

Оба языка, будучи романскими, имеют общие звуковые комбинации, но их частотность и распределение различаются. Испанский язык характеризуется высокой частотностью биграмм с «r» и «a», что может указывать на особенности его фонологии и морфологии. Французский язык отличается высокой частотностью биграмм «er» и «on», что может быть связано с его уникальными грамматическими и лексическими особенностями.

Заключение. В рамках данного исследования проведен сравнительный анализ французского и испанского языков с использованием языка программирования Python и библиотеки pandas. Статистический анализ выявил различия в длине самых длинных слов: «anticonstitutionnellement» (25 символов) во французском и «achicharronariamos» (18 символов) в испанском. При этом средняя длина слов оказалась практически идентичной: 10.09 символов для французского и 10.06 символов для испанского. Частотный анализ биграмм показал, как общие черты, так и различия между языками. Общие биграммы, такие как «ra», «es», «re», «en», «is» и «as», свидетельствуют о схожих звуковых комбинациях и, возможно, общих лингвистических корнях. Однако различия в частотности и наличие уникальных биграмм в каждом языке, например, «ar» и «os» в испанском и «er» и «on» во французском, указывают на специфические особенности фонологии и морфологии каждого языка.

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

COMPARATIVE ANALYSIS OF WORD FREQUENCY AND STRUCTURAL FEATURES OF FRENCH AND SPANISH LANGUAGES 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 Sci., Associate Professor, Dean of the FCAD

Annotation. The article presents a comparative analysis of French and Spanish using Python. The analysis of lexemes, their average length, and bigrams are also examined using the pandas library. Based on the results of the analysis, common patterns for both languages are identified.

Keywords: comparative analysis, lexeme, French, Spanish, average length.