

ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ АНАЛИЗА ДАННЫХ К ИССЛЕДОВАНИЮ ЛЕКСИКИ АНГЛИЙСКОГО ЯЗЫКА

Карлюк П.А., Коршов М.И., Кранцевич В.С.

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

Научный руководитель: Лихачевский Д.В. – к. т. н., доцент, декан ФКП

Аннотация. Проведен статистический анализ словаря английских слов. Произведена визуализация распределения длины слов, частоты начальных и конечных букв. Продемонстрирована ценность применения методов компьютерной лингвистики для анализа языковых данных и лексикографических ресурсов.

Ключевые слова: анализ данных, английский язык, распределение, визуализация.

Введение. Современные задачи обработки естественного языка требуют глубокого понимания структуры и статистики лексических единиц. Несмотря на наличие крупных словарей, лингвистические закономерности, такие как частотность букв редко анализируются количественно. Отсутствие наглядных статистических обобщений затрудняет применение теоретических знаний в практике обработки естественных языков и корпусной лингвистики. В рамках данной работы был проведен статистический анализ обширного списка английских слов. Целью исследования стало выявление закономерностей в распределении слов и букв (начальных и конечных) на основе визуализации и классификации словарных данных.

Основная часть. В качестве данных для исследования был использован обширный словарь английских слов. Данные представляют собой около 370 тысяч уникальных слов. Для обработки данных применялись инструменты *Python*: библиотека *pandas* для расчета статистик, а также средства визуализации *Matplotlib* [1] и *Seaborn* [2]. Первоначально исследуется общая статистика длины слов в выборке. Средняя длина слова составляет около 9,4 символов, медиана – 9 символов, стандартное отклонение – 2,9. Значения указывают на то, что половина слов английского языка имеет длину до 9 символов, а большинство – не длиннее 11 символов. Минимальная длина слова равна 1, в то время как максимальная длина достигает 31 символ. Таким образом, типичная длина английских слов лежит в диапазоне от 7 до 11 символов.

График распределения длины слов в английском языке (рисунок 1):

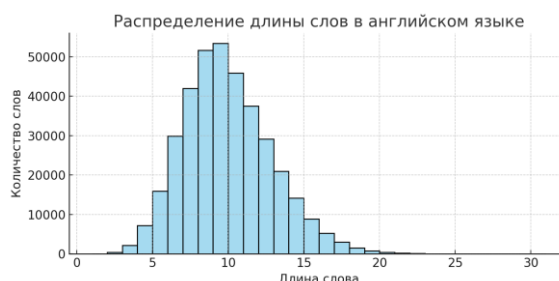


Рисунок 1 – Распределение длины слов в английском языке

Гистограмма показывает, что наиболее часто встречаются слова средней длины. Слова длиной менее 4 или более 15 символов встречаются существенно реже. Данная картина соответствует ожиданиям: большая часть общеупотребительных слов имеет умеренную длину, тогда как очень длинные слова обычно являются либо научно-техническими терминами, либо редкими именами собственными.

Следующим этапом проведен анализ того, с каких символов чаще всего начинаются английские слова. Была выдвинута гипотеза, что распределение начальных символов неравномерно и некоторые символы алфавита значительно доминируют. Для подтверждения гипотезы был построен график частоты начальных символов английских слов (рисунок 2):



Рисунок 2 – Частота начальных символов английском языке

Результаты подсчета частот подтверждают эту гипотезу. Наиболее частый начальный символ – *s*, на него начинаются около 38 761 слова. Высокие показатели также у символов *p*, *c*, *a* и *u*, каждый из которых имеет в начале несколько десятков тысяч слов. Доминирование символа *s* объясняется тем, что в английском языке множество слов начинаются на *s* (*self*-, *sub*-, *super*-), а также тем, что символ *s* входит в состав распространенных корней. Символы *p* (*post*-, *pre*-, *pro*-) и *c* также часто начинают слова, благодаря как заимствованиям из латинского, так и коренным словам. Напротив, ряд символов встречается в начале слов крайне редко. К примеру, на символ *x* начинается всего около 500 слов, *q* – около 1,8 тысяч, *z* – около 1,4 тысяч, что обусловлено фонетическими ограничениями.

Аналогично был проанализирован последний символ слов. Ожидалось, что в конце слов будут чаще встречаться символы, связанные с типичными словообразовательными суффиксами или грамматическими окончаниями. Был построен график частоты конечных символов (рисунок 3):

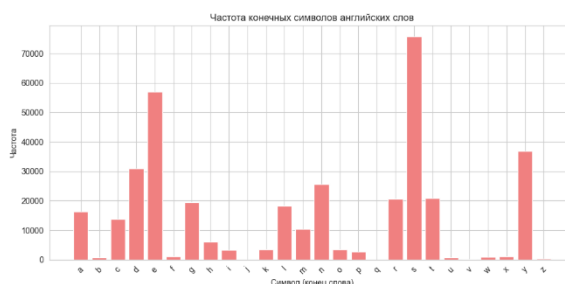


Рисунок 3 – Частота конечных символов английском языке

Данные, представленные на гистограмме, подтверждают эту гипотезу: самый частотный символ в конце английских слов – *s*. Он встречается на конце около 75 тысяч слов, что значительно больше, чем у прочих символов. Это легко объяснимо тем, что *-s* является показателем множественного числа существительных, а также окончанием глаголов в третьем лице единственного числа. Кроме того, высокая частота символа *e* на конце около 57 тысяч слов связана с распространенностью суффиксов на *-e*. Также заметно большое число слов оканчивается на *u* и на *d*. Редко заканчиваются английские слова на символы *j* и *q* в выборке найдено по около 30 слов, в основном заимствования либо сокращения. В целом, частотность конечных символов в значительной степени определяется морфологией: самые распространённые окончания связаны с грамматическими формами или типовыми суффиксами.

Заключение. Проведенный анализ большого списка английских слов позволил выявить ряд закономерностей и подтвердить выдвинутые гипотезы. Во-первых, установлено, что распределение длины слов имеет выраженный максимум на диапазоне 7–11 символов, что отражает типичную структуру английского слова. Во-вторых, частотный анализ символов показал неслучайный характер их использования: наиболее часто слова начинаются на те символы, которые связаны со словообразовательными элементами, а заканчиваются – на символы распространенных грамматических окончаний.

Полученные результаты демонстрируют эффективность применения компьютерного анализа для лингвистических исследований. С помощью программных средств удалось быстро обработать сотни тысяч слов и наглядно представить статистические закономерности.

Список литературы

1. Matplotlib documentation - Matplotlib 3.10.1 documentation [Электронный ресурс]. – Режим доступа: <https://matplotlib.org/stable/index.html>. – Дата доступа: 20.03.2025.
2. Seaborn: statistical data visualization - seaborn 0.13.2 documentation [Электронный ресурс]. – Режим доступа: <https://matplotlib.org/stable/index.html>. – Дата доступа: 20.03.2025.

UDC 004.4

APPLICATION OF DATA ANALYSIS TOOLS TO THE RESEARCH OF ENGLISH VOCABULARY

Karliuk P.A., Karshou M.I., Krantsevich V.S.

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

Likhachevsky D.V. – Cand of Sci., Associate Professor, Dean of FCAD

Annotation. A statistical analysis of English words was carried out. The distribution of word lengths, as well as the frequency of initial and final letters, was visualized. The value of applying computational linguistics methods for analyzing language data and lexicographic resources is demonstrated.

Keywords: data analysis, English, distribution, visualization.