Каршакевіча. – Мінск : Выдавецтва ЦК КПБ, 1988. – 47 с. : іл. – (Бібліятэка “Вожыка”; № 1 (187))

5. Сказ пра Лысую гару : паэма / Францішак Вядзьмак-Лысагорскі; [прадмова В. Блакіт]; мастацкае афармленне Ю. Л. Рыжыкава. – 2-е, поўнае выд. – Мінск : Выдавецтва ЦК КПБ, 1991. – 63 с. : каляр. іл. – (Бібліятэка “Вожыка”; № 4 (208))

6. З Францішкам Ведзьмаком-Лысагорскім гутарыць “Крыніца” // Крыніца. – 1995. – № 1–2. – С. 46–48.

7. Нил Гилевич: "Пісаў як пісалася" // Вечерний Минск. – 2003. – № 212. – С. 3

УДК 81’322.2

Студ. А.В. Кочемарова
Науч. рук. канд. филол. наук, доц. каф. А.А. Акушевич
(кафедра редакционно-издательских технологий, БГТУ)

АТРИБУЦИОННЫЕ ВОЗМОЖНОСТИ ПРОГРАММЫ «АНАЛИЗАТОР ТЕКСТА» (НА ПРИМЕРЕ ТЕКСТОВ УСТАНОВЛЕННОГО АВТОРСТВА)

В свободном доступе недостаточно программ, которые могли бы использоваться для определения авторства. Наиболее востребованные сервисы: Атрибутор, Лингвоанализатор (для русскоязычных художественных произведений) и ИС Смалт (для публицистики 60-70 гг. 19 века) [1] – позволяют сравнить загруженный пользователем текст только с имеющимися в базе примерами. Таким образом, возникает потребность в программах, которые бы всесторонне анализировали текст, позволяли на основе полученных статистических данных выявлять черты авторского стиля и/или делать выводы о возможном авторстве. Однако основной проблемой формального (машинного) анализа текста является отсутствие различающей способности, т.е. близкие значения для большинства авторов, и отсутствие устойчивых показателей [2].

Программа «Анализатор текста» является разработкой американского студента Юнуса Кулиева в рамках учебного процесса. Она предназначена для оценки эссе, но обладает широким функционалом, включающим в себя:

– подсчёт количества знаков, слов, предложений, абзацев;

– выведение коэффициента разнообразия слов и сложности восприятия текста;

– определение скорости прочтения, поиск самого длинного слова и предложения в тексте;

– анализ текста, представленного в виде фотографии, изображения;

– составление рейтинга из чаще всего встречающихся слов с указанием их количества в тексте и процентным отношением к общему объёму текста;

– отображение в виде таблицы и графика распределения слов по количеству их знаков;

– синтаксический анализ, анализ объектов текста (вывод ссылок на статьи сайта Википедия для ключевых терминов и имён собственных), анализ эмоциональной окраски текста, формирование рекомендаций по улучшению [3].

Нами был проведен эксперимент с использованием данного приложения. Функции, указанные выше последним пунктом, на тот момент не работали. Приложением «Анализатор текста» в процессе исследования были проверены отрывки длинной в 4000 слов из трёх разных произведений каждого из трёх выбранных авторов. Писатели и произведения подбирались с учётом небольшого временного промежутка (конец 19 – начало 20 века), чтобы избежать стилистических различий, связанных с особенностями языка разных эпох.

Первой группой стали произведения А. П. Чехова «Человек в футляре», «Учитель словесности» и «Анна на шее». Соотношение разнообразия слов составило 48,45%, а показатель читаемости – 259,73.

Второй группой были романы М. А. Булгакова «Белая гвардия», «Роковые яйца» и «Записки юного врача». Соотношение разнообразия слов в тексте – 51,95 %, читаемость – 836,6.

Третья группа состояла из произведений И. А. Бунина «Суходол», «Митина любовь», «Жизнь Арсеньева». Соотношение разнообразия слов – 50,02 %, показатель читаемости – 535,6.

Данные по соотношению разнообразия слов у трёх авторов не сильно различаются, что вызывает сомнения в верности результатов. А коэффициент читаемости, определяемый по средней длине предложений и количеству запятых в них, имеет напротив чёткое разграничение. Чем больше этот коэффициент, тем сложнее текст для восприятия.

График распределения слов по их длине (рис. 1) показывает сильное различие в употреблении авторами слов от 2 до 6 символов. Тексты Михаила Булгакова содержат заметно меньше слов длиной от 2 до 5 символов, однако чуть больше слов от 7 до 11 символов. Начиная с показателей относительно слов от 7 символов, кривые для А. П. Чехова и И. А. Бунина идут близко друг к другу. Произведения

И. А. Бунина имеют наиболее сглаженную кривую на отрезке слов от 2 до 5 символов.

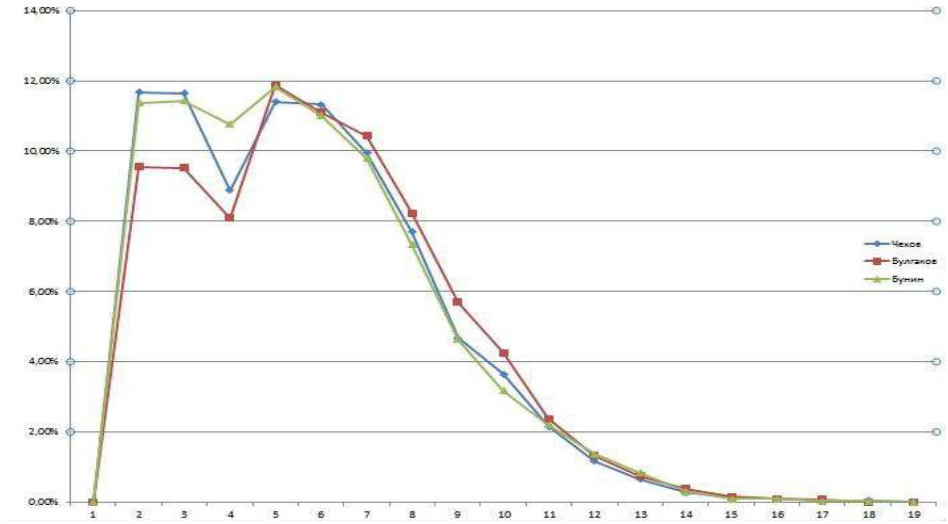

Рисунок 1 – Схема распределения слов по количеству знаков

Результаты, представленные в качестве рейтинга из наиболее встречающихся в тексте слов, нельзя назвать удовлетворительными, так как одно слово в двух разных его формах (например, падежных) приложение подсчитывает как две отдельные текстовые единицы. Показатель разнообразия слов в тексте при этом определяется верно, что было проверено с помощью сайта «Анализатор текста» [4].

Исходя из полученных данных, можно сказать, что приложение «Анализатор текста» для задач установления авторства слабо подходит, хоть и может быть полезно для быстрого получения некоторых статистических сведений (разнообразие слов, восприятие текста, построение кривой и таблицы употребления слов разной длины).

ЛИТЕРАТУРА

1. Статистические методы анализа литературного текста / Komiwiki [Электронный ресурс]. – 2012. – Режим доступа: http://komiwiki.syktsu.ru/index.php/%. – Дата доступа: 10.04.2021.

2. Батура, Т. В. Формальные методы определения авторства текста. Т 10, вып. 4 / Т. В. Батура. – Новосибирск: Вестник НГУ, 2012. – 84 с.

3. TextAnalyzer Pro / Playмаркет [Электронный ресурс] – Режим доступа: https://play.google.com/store/apps/details?id=com.graspery.www.wordcountpro. – Дата доступа: 10.04.2021.

4. Анализатор текста TextAnalyzer / Devaka [Электронный ресурс]. – Режим доступа: https://www.textanalyzer.ru. – Дата доступа: 11.04.2021.