DOI: 10.33184/dokbsu-2020.1.10

Обработка лингвистических ресурсов на языке программирования Питон (на материале произведений Р. У. Эмерсона, Э. Дикинсон, Э. По)

Е. М. Черепанова*, Р. Г. Мифтахова

Башкирский государственный университет
Россия, Республика Башкортостан, 450076 г. Уфа, улица Заки Валиди, 32.

**Email: cherepanovaem@mail.ru*

Романтизм в американской литературе является одним из сильнейших интеллектуальных направлений. Р. У. Эмерсон, Э. Дикинсон, Э. А. По – авторы, уделившие пристальное внимание природе, ее генезису, взаимодействию человека и природы. Литературные критики обращают внимание на разницу в понимании писателями природы и ее влияния на человека. Работы Р. Эмерсона, Э. Дикинсон и Э. По были использованы для изучения способов обработки литературных текстов на языке программирования Python. Python измерил лексическое разнообразие в текстах авторов, определил семантическое поле слова «природа» в каждом тексте и оценил количество информативных и неинформативных слов. Оказалось, что стихи Э. Дикинсона имеют самый высокий показатель лексического разнообразия и наименьшее количество неинформативных слов. Компьютерный анализ подтвердил результаты традиционных методов лингвистического анализа. В статье рассматриваются способы обработки литературных текстов на языке программирования Python. Результаты литературоведческого анализа произведений Эмерсона, Дикинсон и По, выполненного на основе традиционных методик, совпали с результатами, полученными с помощью компьютерной обработки текстов.

Ключевые слова: романтизм, коэффициент лексического разнообразия, семантическое поле, информативные слова, неинформативные слова.

Ральф Уолдо Эмерсон, Эдгар Аллан По, Эмили Дикинсон – выдающиеся представители американского романтизма. В их произведениях наиболее полно отразились особенности романтической литературы Нового Света с ее пристальным вниманием к вопросам познания окружающего мира, его происхождения, роли божественного начала в обустройстве вещного мира, природы как гармоничного начала, благотворно влияющего на человека и его душу.

С исторической точки зрения интерес американских романтиков к природе можно объяснить тем, что на их сознание, в значительной мере, повлиял тот факт, что к концу

XVIII столетия, одержав победу в Войне за Независимость, американцы в полной мере почувствовали себя уже не колонистами, но хозяевами Северной Америки. Первые колонии, как известно, располагались вдоль восточно-атлантического побережья, в то время как огромные территории в центре и на западе континента оказались совершенно не освоены. Прерии, белоснежные вершины гор, величественные каньоны, стремительные реки и непроходимые леса будоражили сознание американцев. Так в американском романтизме возникает фронтир – уникальное литературное направление, описывающее освоение американцами необжитых территорий. Самым ярким литературным персонажем фронтира стал Натти Бампо – герой историко-приключенческой пентологии Ф. Купера.

Следует отметить, что, несмотря на всеобщий интерес к феномену природы, американские романтики по-разному трактовали как само понятие природа, так и ее происхождение, а также влияние природы на человека.

Ральф Уолдо Эмерсон рассматривает природу, как нечто безусловно прекрасное и гармоничное, способное сделать более совершенным и самого человека. По Эмерсону, природа великолепна в каждом своем проявлении. Близость к природе способствует восстановлению не только физических, но и душевных сил человека.

Эмили Дикинсон – американская поэтесса, чье творчество считается вершиной романтической поэзии США. Большая часть ее стихотворных произведений посвящена природе в ее каждодневном, бытовом проявлении. У Дикинсон природа – это утро и вечер, закат и рассвет, это пение соловья или ветер, как гость, заглянувший в ее сад.

Эдгар По – писатель, поэт рассматривает природу более глобально. Его произведение «Эврика» – это космогоническое сочинение, в котором автор приводит свою версию возникновения Вселенной. Не будучи ученым, По интуитивно предугадал многие важнейшие открытия, совершенные в XIX и XX веках в области астрономии, математики и физики, такие как пульсирующая Вселенная, первичное вещество (единица), гравитация, концепция Большого Взрыва. По существу По предвосхитил ньютоновскую теорию Вселенной.

Практическим материалом для обработки лингвистических ресурсов на языке программирования Python стали эссе Эмерсона «Природа», стихотворения Дикинсон и поэма в прозе По «Эврика».

Поскольку тематически произведения Эмерсона, Дикинсон и По близки, то первая задача состояла в том, чтобы определить коэффициент лексического разнообразия произведений с помощью языка Python. Под коэффициентом лексического разнообразия (lexical diversity) в программировании понимается количественная характеристика текста, отражающая степень богатства словаря, использованного в тексте. Основным

показателем лексического разнообразия принято считать соотношение числа отдельных лексических единиц и количества их употреблений в тексте.

Для решения данной задачи тексты были разбиты на лексические единицы. Затем для более точного подсчета вхождений лексических единиц слова, написанные с прописной буквы, были заменены на слова, написанные со строчной буквы. Далее был создан и использован применительно к каждому тексту код определения коэффициента лексического разнообразия. Произведение Р. Эмерсона получило коэффициент 4.77, тексты Э. Дикинсон и Э. По получили соответственно 4.66 и 9.422. Коэффициент лексического разнообразия показывает сколько раз в среднем каждое слово встречается в тексте, а значит, чем меньше этот коэффициент, тем текст лексически разнообразнее

Полученные данные показали, что произведения Э. Дикинсон обладают более низким коэффициентом лексического разнообразия, то есть, на единицу объема текста приходится максимальное количество употреблений лексем. Поскольку в исследуемых текстах наиболее часто употребляемым является слово “nature”, то следующий этап анализа текстов заключался в том, чтобы определить его семантическое поле. Как известно, семантическое поле – это смысловая парадигма, в которую входят различные слова, имеющие общий семантический признак. Слово “nature” может считаться доминантой во всех проанализированных текстах, но семантическое поле у каждого автора индивидуально.

В программе Питон оно определялось по контекстуальному признаку. Так, в одном контексте со словом ’nature’ у Р. Эмерсона встречаются такие слова как man, god, religion, wisdom, truth и прочие, а для произведений Э. По система выдала space, stars, gravitation, sphere и другие.

Полученные результаты показали, что у Эмерсона в семантическое поле слова «природа» входят такие слова как человек, красота, сила, жизнь, наука, мудрость и др. У По семантическое поле слова «природа» включает космос, расстояние, звезды, скопления, гравитация, бог. В поэтических текстах Дикинсон определить семантическое поле слова «природа» не удалось, поскольку машина не может дать точный анализ, если объем анализируемого текста относительно небольшой. Однако известно, что поэт ассоциирует природу с такими понятиями как бог, небеса, незнакомка, красота, гармония.

Следовательно, результаты машинного анализа подтверждают, что для Эмерсона природа – источник жизни, она раскрывает себя в своей красоте, природа познаваема и является– источником истины и мудрости, природа связана с человеком. Для Дикинсон природа – проявление бога, она гармонична, красива, но для поэта она незнакомка, то есть, непознаваема. В понимании По природа – есть космос, скопление звезд, гравитация, расстояние.

Лингвистическая обработка на языке программирования также предоставляет возможность сравнить тексты с помощью количественных показателей, например, определить соотношение информативных и неинформативных слов. Информативными словами принято называть значимые, общеупотребительные слова, которые позволяют определить тематику и жанр текста. К таким словам можно отнести, например, “nature”, “reason, science”, “universe”. Для определения соотношения информативных и неинформативных слов в анализируемых текстах был составлен еще один код. При проведении исследования был использован корпус стоп-слов Natural language ToolKit (NLTK). В результате были получены следующие данные:

В тексте У. Р. Эмерсона информативные слова составили 48%, в текстах Э. Дикинсон - 53%, а в текстах Э. По – 46%

Анализ показал, что в произведениях Дикинсон количество неинформативных слов наименьшее, что можно объяснить высоким коэффициентом лексического разнообразия в ее произведениях.

Таким образом, анализ художественных текстов на языке программирования Python позволил определить их коэффициент лексического разнообразия, семантическое поле слова “nature” у каждого из авторов и сравнить соотношение информативных и неинформативных слов в текстах Эмерсона, Дикинсон и По. Такой подход позволил расширить понимание особенностей творчества авторов с лингвистической точки зрения. Полученные результаты подтверждают выводы, сделанные на основе литературоведческого анализа художественных текстов. Использование современных систем компьютерного анализа текстов позволяет решать лингвистические задачи значительно быстрее, чем с помощью традиционных литературоведческих приемов и методик.

Литература

1. Дикинсон Э. Стихотворения (Предисл. и пер. А. Гаврилова). – М.: Радуга, 2001. – 441 с.
2. Ковалев Ю. В. Эдгар Аллан По. Новеллист и поэт: Монография. –Л.: Худож. лит., 1984. 296 с.
3. Морозкина Е. А. Мифтахова Р. Г. Влияние информационных технологий на развитие лингвистических норм. Вестник БашГУ. 2012. №1.– С. 162–164.
4. Осипова Э. Ф. Загадки Эдгара По. Исследования и комментарии. – СПб.: Филологический факультет СПбГУ, 2004. – 172 с.
5. Осипова Э. Ф. Ральф Эмерсон и американский романтизм. СПб. : Изд-во С.-Петерб. ун-та, 2001. – 190с.
6. Beaver, H. The Science Fiction of Edgar Allan Poe. London: Penguin Books, 1976.-466p.
7. Cody, John. After Great Pain: The Inner Life of Emily Dickinson. – Cambridge, MA: Harvard University Press, 1971. – 195p.
8. Dickinson E, Poems. – NY: Bookworld, 1999. – 258p.

9. Farr, Judith. The Passion of Emily Dickinson. – Cambridge: Harvard UP, 1992. – 439p.
10. Paul Sh. Emerson's Angle of Vision: Man and Nature in American Experience. Cambridge, Mass., 1969, p. 230.

Статья рекомендована к печати кафедрой лингводидактики и переводоведения БашГУ
(докт. филол. наук, проф. Морозкина Е. А.)

Processing of linguistic resources in Python programming language (based on the works of R. W. Emerson, E. Dickinson. E. A. Poe)

E. M. Cherepanova*, R. G. Miftakhova

Bashkir State University
32 Zaki Validi Street, 450076 Ufa, Republic of Bashkortostan, Russia.

**Email: cherepanova.em@mail.ru*

Romanticism in American literature is one of the strongest intellectual currents. R. W. Emerson, E. Dickinson, E. A. Poe are the writers who paid keen attention to nature, its genesis, the interaction between man and nature. Literary critics dwell on difference of the writers' comprehension of nature and its influence on man. The works of R. Emerson, E. Dickinson and E. Poe were used to study ways of literary texts processing in Python programming language. Python measured lexical diversity in the texts of the authors, determined the semantic field of the word "nature" in each text and rated the number of informative and non-informative words. It turned out that E. Dickinson's poems have the highest index of lexical diversity and the least number of non-informative words. The computer analysis proved the results of traditional methods of linguistic analysis.

Keywords: Romanticism, lexical diversity, semantic field, informative words, non-informative words.