

Спицына Наталья Александровна

канд. филол. наук, доцент

Ходаницкая Владислава Викторовна

студентка

ФГАОУ ВО «Дальневосточный федеральный университет»

г. Владивосток, Приморский край

DOI 10.31483/r-74236

КОРПУСНЫЕ ТЕХНОЛОГИИ КАК МЕТОД ЦИФРОВОЙ ГУМАНИТАРИСТИКИ (ОПЫТ АНАЛИЗА ПРЕДВЫБОРНОГО ДИСКУРСА)

***Аннотация:** в статье рассматриваются информационные возможности корпусной лингвистики на материале предвыборного дискурса. Исследование основано на применении корпусных менеджеров AntConc и SketchEngine. Применение инструментов корпусной лингвистики является примером изучения фактов естественного языка с применением методов компьютерной лингвистики и информационных технологий. В статье выделены списки наиболее частотных лексем предвыборного дискурса Б. Обамы и Д. Трампа, показаны коллокационные профили дискурсов предвыборной кампании.*

***Ключевые слова:** цифровая гуманитаристика, информационные технологии, корпусная лингвистика, корпусный менеджер, лингвистическая корпус, частотность, предвыборный политический дискурс.*

Одним из стремительно развивающихся направлений гуманитарных наук в последнее десятилетие стала «цифровая гуманитаристика» («цифровые гуманитарные науки», «Digital Humanities»), которая представляет собой междисциплинарную область исследований. Она объединяет методики и практики гуманитарных, социальных и вычислительных наук с целью изучения возможностей применения и интерпретации новых цифровых и информационно-коммуникационных технологий, систематического использования цифровых ресурсов в гуманитарных и социокультурных исследованиях и образовании [2]. Это междисциплинарное направление в развитии гуманитарных наук использует методы, понятия

и технологии информатики. Корпусная лингвистика входит в область цифровой гуманитаристики. В широком понимании составление и использование корпусов текстов является эмпирическим текстоориентированным методом исследования языкового материала с применением информационных технологий и компьютерной лингвистики. Основой корпусной лингвистики является разработка теоретических основ и практических приемов построения, машинной обработки и эксплуатации языковых данных, оформленных как корпус текстов. Такие корпуса составляют машинописное, стандартно организованное собрание репрезентативных для определенного языка диалектов или другого подмножества языков письменных или устных текстов в электронной форме, предназначенных для лингвистического анализа и описания. Это своего рода информационно-справочная система.

В современном языкознании лингвистические корпуса широко используются для изучения особенностей и специфики употребления в речи единиц различных языковых уровней.

Стоит отметить, что применение методов и программного обеспечения отрасли корпусной лингвистики дает дополнительные возможности для изучения политического дискурса, в частности его количественных показателей, и таким образом делать выводы об определенных направлениях и тенденциях коммуникативной стратегии участников политических процессов. В работе был произведен анализ политического дискурса на основе статей, посвященных предвыборному процессу двух кандидатов – Барака Обамы, Дональда Трампа.

В исследовании используется CD (corpus-driven) подход [1]. На основе 102 статей на английском языке, посвященных предвыборному процессу в Соединенных Штатах Америки, был составлен небольшой корпус текстов в 500,000 слов. Для исследования был использован корпус-менеджер AntConc версии 3.5.8. Это мультиплатформенная программа для проведения лингвистических исследований и управления данными корпусов, разработанная Энтони Лоуренсом в Японии, Waseda University [3]. С помощью данного корпус-менеджера

был составлен тезаурус из 20 самых часто используемых ключевых слов (seedwords).

Все 20 ключевых слов разделены на группы:

- 1) глаголы: to cast, to appeal, to vote;
- 2) слова, относящиеся к внешней и внутренней политике государства: Internal, foreign, domestic, strategy, policy, actions;
- 3) слова, относящиеся к процессу политических выборов: elections, primaries, candidates, ballot, electioneering, referendum, electorate, choice;
- 4) слова – обращения к электорату: respected, believe, dear;

Для дальнейшей работы использовался другой корпус-менеджер Sketchengine, разработанный в компании Lexical Computing Limited в 2003 году [4]. Его главное отличие от корпус-менеджера AntConc состоит в том, что в его систему управления входит инструмент WebBootCat, который с помощью ключевых слов (seedwords), автоматически создает нелингвистический неаннотированный корпус.

С помощью данного инструмента был создан корпус, который называется «Corpus methods of research of pre-election discourse», состоящий из 353 статей на английском языке, посвященным предвыборному процессу в Соединенных Штатах Америки. Статьи отобраны по периоду времени, датируемому с 2010 по 2017 года.

Хотелось бы отметить преимущество автоматического поиска корпус-менеджера Sketchengine, это то, что источниками большинства статей являются официальные проверенные источники, а не жёлтая пресса.

Далее корпус «Corpus methods of research of pre-election discourse» был переведен в формат «ТХТ», что позволило дальше использовать AntConc. С помощью функции Wordlist и заранее отобранного stoplist, состоящего из предлогов, артиклей, частиц и союзов, а так же вводных слов, автоматической выборкой было отобрано 50 самых часто-используемых слов в материале данного корпуса.

Показатели частности

Позиция в списке	Слово	Частота использования
- 33	election	2811
- 45	strategy	2080
- 46	company	2063
- 50	vote	1974
- 53	domestic	1848
- 58	people	1698
- 59	business	1673
- 61	party	1645
- 64	foreign	1538
10–67	states	1451
11–69	violence	1420
12–73	president	1326
13–79	policy	1253
14–81	government	1247
15–88	court	1194
16–92	time	1172
17–94	political	1155
18–95	market	1150
19–99	national	1121
20–102	competitive	1083
21–109	law	1008
22–111	support	987
23–114	candidates	983
24–117	industry	969
25–118	world	962
26–120	campaign	957
27–122	public	950
28–125	each	934
29–128	electoral	904
30–130	united	896
31–137	process	839
32–144	work	751
33–146	service	748
34–148	international	742
35–152	system	731
36–157	number	709
37–160	change	705

38–161	local	704
39–170	ballot	662
40–171	general	662
41–173	value	659
42–177	act	652
43–185	power	635
44–203	issues	580
45–214	right	562
46–265	security	465
47–270	countries	452
48–272	major	451
49–273	country	448
50–134	poll	446

Таблица разделена на три колонки – позиция слова (по порядку и в корпусе), само слово и частота использования слова в материале корпуса. Можно сделать некоторый анализ темы дискурса, то есть его основного направления, используемого в данном корпусе. Из слов, которые имеют более 1000 словоупотреблений в корпусе, 5 относятся к предвыборному процессу, 10 к внешней политике государства и его внутренней системе управления, 3 направлены на рыночную торговлю и бизнес, 3 на законодательную систему управления.

Частотные слова включают в себя как терминологическую лексику, например, ballot, electoral, campaign, strategy, government, poll, так и нетерминологическую, например, time, each, right, power. Количественные показатели частотности употребления терминологической и нетерминологической лексики позволяют выделить темы, преобладающие в политическом дискурсе кандидатов. Частотность дает представление о том, какое место отводил каждый из говорящих в своих речах терминологии соответствующей отрасли, какая из них преобладала в его риторике, какую часть от общего количества слов у каждого из кандидатов занимала специальная лексика и, наконец, какую часть политического вещания занимала терминология в целом.

Для выполнения данной задачи с помощью корпус – менеджера AntConc все тексты статей в корпусе на основе самых частотных слов были разделены на две группы, то есть впоследствии два подкорпуса. Это материалы статей,

посвященных предвыборному процессу кандидатов в президенты в Соединенных Штатах Америки Барака Обамы и Дональда Трампа, так как именно этих материалов больше всего используется в составленном общем корпусе «Corpus methods of research of pre-election discourse». В приведенных ниже таблицах слова не только разделяются по адресантам, но и произведен анализ кластерных коллокатов и частота их употребления.

Таблица 2

Коллокация в предвыборных речах Б. Обамы

№	Word	№	Clusters	Frequency
2	Strategy	3	execution	82
		6	Strategy-making	46
		11	Strategy-critical	30
		12	Strategy-supportive	26
3	Company	2	resources	27
		4	direction	25
		8	situation	19
		9	managers	15
8	Party	3	system	37
		4	affiliation	31
		6	candidate	29
		7	leader	28
10	State	1	licensing	162
		5	council	89
		6	legislature	78
		8	department	75
20	Competitive	1	advantage	222
		2	capabilities	94
		3	strength	90
		4	position	81
21	Law	1	enforcement	100
		3	commission	97
		6	project	91
		7	act	70
27	Public	1	housing	53
		2	health	43
		3	opinion	41
		4	assistance	37
31	Country	2	markets	41
		6	differences	36

		7	Country-specific	23
		9	coordination	17
32	Process	3	can	46
		5	would	37
		6	has	34
		9	may	21
33	Work	1	together	31
		2	closely	30
		6	environment	27
		8	better	20
34	Service	1	providers	52
		6	industries	43
		7	Commission	40
		10	provision	31
36	Systematic	1	analysis	30
		2	behavior	23
		3	attempt	22
		4	implementation	20
44	Act	1	Provides	74
		4	applies	64
		5	allows	34
		6	requires	30
45	Power	2	tools	35
		3	equipment	34
		4	systems	23
		5	absolute	20
46	Issues	3	related	30
		5	continue	30
		6	concerning	28
		7	Including	24

Таблица разделена на пять колонок:

- 1) позиция слова в предыдущей таблице в зависимости от частоты его использования в материале корпуса;
- 2) само слово;
- 3) позиция кластерного коллоката в корпусе (по частоте использования);
- 4) кластерный коллокат (первые 4, которые являются самыми значимыми и самыми частотными);
- 5) частота использования кластера в материале корпуса.

Коллокация в предвыборных речах Дональда Трампа

№	Word	№	Clusters	Frequency
5	Domestic	1	violence	153
		2	abuse	65
		3	incident	59
		4	investors	45
6	People	1	who	126
		5	have	89
		8	can	71
		9	will	61
		11	should	53
7	Business	1	units	80
		3	review	47
		5	strategy	43
		8	model	41
9	Foreign	1	policy	329
		2	investors	144
		3	investment	139
		4	markets	100
14	Government	3	policy	115
		5	approval	112
		6	officials	112
		8	agencies	89
18	Market	1	share	103
		2	access	87
		3	position	87
		4	conditions	66
24	Industry	3	regulator	50
		4	attractiveness	40
		5	value	33
		7	leaders	30
37	Number	2	one	51
		4	two	50
		5	changes	37
		6	allowed	34
3	Change	3	their	36
		4	your	34
		5	that	29
		7	our	28
39	Local	1	government	31

		2	domestic	28
		8	statement	44
		9	issues	40
43	Value	1	chain	265
		2	added	96
		6	activities	81
		7	judgment	80
47	Right	3	now	39
		4	thing	34
		5	Hand	34
		7	decision	30

В предвыборной риторике Барака Обамы чаще всего используются слова и кластерные коллокации, связанные с государством, например party, state, country, и с государственным управлением, связанным с народовластием, например, strategy, party leader, law enforcement, public health, work together, service, power tools, systematic. Больше всего в тематике данного адресанта используются слова, связанные с оценочностью и эффективностью, причем в равной степени. Из этого следует, что в данном материале дискурс Барака Обамы направлен на идеологизацию некоторых политических идей, а также внушение адресатам политически «правильной» позиции.

В предвыборном дискурсе Дональда Трампа явно прослеживается тенденция употребления слов и кластерных коллокаций, отнесенных по тематике к международной торговле и товарообороту, например, такие слова как foreign investors, business strategy, market conditions, industry attractiveness, value. Так же содержатся слова и выражения, призывающие электорат к изменениям, например, people can, change your, change our, right now.

Из количественного анализа слов, разделенных по характеристикам, следует, что он в большей мере использует слова, направленные на эффективность, то есть убедить электорат в своей позиции и побудить к действию. В меньшей мере представлены показатели оценочности и отстаивания точки зрения, что говорит о том, он не ставит своей целью привлечь слушателей на свою сторону и

объяснить, что данная позиция лежит в их интересах, но показывает «правильную» позицию и сразу побуждает к действию.

Таким образом, результаты анализа, осуществленного с применением корпусных методов исследования, показали, что политический дискурс Барака Обамы направлен на государственное управление и народовластие, а Дональд Трамп ориентирован на бизнес, международную торговлю и товарооборот.

Список литературы

1. Горина О.Г. Инструменты корпусного анализа в обучении иностранному языку // Вестник Томского государственного университета. – 2018. – №435 [Электронный ресурс]. – Режим доступа: <https://cyberleninka.ru/article/n/instrumenty-korpusnogo-analiza-v-obuchenii-inostrannomu-yazyku> (дата обращения: 23.12.2019).

2. Можаяева Г.В. Цифровая гуманитаристика: организационные формы и инфраструктура исследований / Г.В. Можаяева, П.Н. Можаяева-Ренья, В.А. Сербин // Вестник Томского государственного университета. – 2014. – №389 [Электронный ресурс]. – Режим доступа: <https://cyberleninka.ru/article/n/tsifrovaya-gumanitaristika-organizatsionnye-formy-i-infrastruktura-issledovaniy-1> (дата обращения: 27.12.2019).

3. Anthony L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. URL: <https://www.laurenceanthony.net/software>

4. Kilgarriff A., Rychlý P., Smrž P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. Lorient, 2004. P. 105–116.