

ПАРУЛИНА ИРИНА ЮРЬЕВНА

*Доцент кафедры лингвистики и перевода,
научный сотрудник Управления научной
и инновационной деятельности
Южно-Уральского государственного университета
(национального исследовательского университета)
parulinai@susu.ru*

СОЛОПОВА ОЛЬГА АЛЕКСАНДРОВНА

*Профессор кафедры лингвистики и перевода,
старший научный сотрудник Управления научной
и инновационной деятельности
Южно-Уральского государственного университета
(национального исследовательского университета)
solopovaoa@susu.ru*

КОШКАРОВА НАТАЛЬЯ НИКОЛАЕВНА

*Профессор кафедры международных отношений,
политологии и регионоведения,
старший научный сотрудник Управления
научной и инновационной деятельности
Южно-Уральского государственного университета
(национального исследовательского университета)
koshkarovann@susu.ru*

МЕДИАДИСКУРСЫ СТРАН БРИКС: КОРПУСНЫЕ И КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ИССЛЕДОВАНИИ ЯЗЫКА¹

Аннотация. Статья посвящена анализу использования корпусных и компьютерных технологий в исследовании медиадискурса стран БРИКС. Представлены результаты использования следующих программ: Sketch Engine, AntConc, LanaKey и Voyant Tools. Описанные инструменты дают возможность в визуальной форме презентовать содержательную и тематическую составляющую большого массива данных. Среди главных преимуществ апробированных инструментов отмечается их сбалансированность, наличие разметки, возможность получения точных статистических данных. С учетом перечисленных факторов лингвист может более точно и корректно интерпретировать результаты на фоне широкого экстралингвистического контекста, что особенно важно в контексте деятельности международных объединений. Применение указанных корпусных инструментов демонстрирует тенденцию цифрового вектора развития лингвистики, когда совмещается интеллектуальный анализ данных и визуализация данных.

Ключевые слова: медиадискурс, корпусные технологии, компьютерная обработка текста, БРИКС.

PARULINA IRINA YURIEVNA

*Associate Professor at the Department of Linguistics and Translation,
Researcher at Research and Innovation Services
South Ural State University (national research university)*

SOLOPOVA OLGA ALEXANDROVNA

*Professor at the Department of Linguistics and Translation, Senior
Researcher at Research and Innovation Services
South Ural State University (national research university)*

KOSHKAROVA NATALYA NIKOLAEVNA

*Professor at the Department of International Relations, Political Science and Regional Studies,
Senior Researcher at Research and Innovation Services
South Ural State University (national research university)*

¹ Исследование выполнено за счет гранта Российского научного фонда № 24-18-00049. URL: <https://rscf.ru/project/24-18-00049/>

MEDIA DISCOURSES OF THE BRICS COUNTRIES: CORPUS AND COMPUTER TECHNOLOGIES IN LANGUAGE RESEARCH¹

Abstract. The paper deals with the use of corpus and computer-assisted technologies in the process of analysing BRICS countries' media discourse. The authors provide an overview of the following applications used to process the corpora: Sketch Engine, AntConc, LanaKey, and Voyant Tools. The tools give an opportunity to present content and thematic characteristics of big data. Other advantages include their equation, annotation, and the opportunity to obtain exact statistic results. These merits of the tools under analysis let the scholar precisely interpret the findings amid the extra-linguistic context which is of great importance in the context of the international organisations' activities. The use of the corpus technologies exhibit the linguistic trend to combine intellectual analysis and the data visualisation.

Key words: media discourse, corpus technologies, text processing techniques, BRICS.

В эпоху быстрого роста информационных технологий и глобализации, медиадискурс является ключевым фактором, влияющим на социальные, политические и культурные процессы в современном мире. Изучение медиадискурса приобретает все большее значение для филологов и лингвистов в связи с ростом влияния медиа на языковые процессы и общественное сознание. Настоящее исследование направлено на анализ использования корпусных и компьютерных технологий в лингвистике медиадискурсов стран БРИКС (Бразилия, Россия, Индия, Китай, ЮАР).

Исследование медиадискурса в странах БРИКС представляет собой сложную задачу, учитывая разнообразие языков (португальский, русский, хинди, китайский, английский), культурных особенностей и политических систем этих регионов. Традиционные методы анализа могут оказаться недостаточно эффективными и приводить к неполным или неточным результатам. В этом контексте, компьютерные и корпусные технологии становятся неотъемлемой частью современного исследования медиадискурса, обеспечивая более точный и объемный анализ данных. Основные причины, обосновывающие необходимость использования этих технологий, включают возможность обработки больших объемов данных и обеспечение объективности (компьютерные алгоритмы обработки данных основаны на четких математических моделях и статистических методах, что позволяет достичь более надежных и объективных результатов).

Корпусные технологии оптимизируют процесс исследования медиадискурса, позволяя собирать, организовывать и анализировать большие коллекции текстовых данных. Для исследования медиадискурсов стран БРИКС была осуществлена компиляция корпусов медиатекстов бразильских, российских, индийских, китайских и южноафриканских СМИ. Для анализа корпусов были применены инструменты корпусного анализа Sketch Engine, AntConc, LanaKey и Voyant Tools. Представим некоторые результаты, полученные при обработке указанными инструментами англоязычного корпуса ЮАР, общим объемом 170 полнотекстовых статей (более 210 тысяч слов), и русскоязычного корпуса, общим объемом 244 полнотекстовые статьи (более 30 тысяч слов).

Sketch Engine² — программное обеспечение для анализа текста, разработанное британской компанией Lexical Computing, одной из ведущих в области обработки естественного языка. Основным назначением инструмента является осуществление анализа и поиска в больших текстовых корпусах в соответствии запросом исследователя. В настоящее время инструмент доступен для работы более чем на 90 языках. Для анализа медиадискурсов стран БРИКС были использованы следующие функции: Word sketch, Concordance, Wordlist, Keywords. Word sketch обрабатывает коллокации заданного слова, группируя результаты по грамматическим отношениям и частям речи (рис. 1).

¹ The research is financially supported by Russian Science Foundation № 24-18-00049. URL: <https://rscf.ru/project/24-18-00049/>

² Sketch Engine. URL: www.sketchengine.eu (дата обращения: 11.09.2024).

modifiers of "Brics"	nouns modified by "Brics"	verbs with "Brics" as object	verbs with "Brics" as subject
five-member ... five-member Brics	summit ... the Brics summit	join ... applied to join Brics	position ... Brics is positioning
reason ... reason Brics	currency ... a common Brics currency	attend ... to attend the Brics (Brazil	stand ... Brics stands
side ... side Brics	country ... the Brics countries	salvage ... salvage the Brics	condemn ... Brics had not condemned
economy ... economies , Brics	bloc ... in joining the Brics bloc	cast ... Brics is often cast	take ... Brics takes
strong ... strong Brics	SA ... a member of Brics , SA has been approached	criticise ... criticising Brics	be ... Brics is

Рис. 1. Коллокации лексемы BRICS в корпусе ЮАР (Sketch Engine, Word sketch)

Concordance — инструмент с различными вариантами поиска. Он выполняет поиск по словам, словосочетаниям, тегам и т.д. и отображает результаты в контексте в виде соответствия (KWIC (key words in context) — ключевые слова в контексте). Результаты могут быть отсортированы, отфильтрованы по ряду критериев и обработаны дополнительно для получения желаемого результата (рис. 2).

	Details	Left context	KWIC	Right context
1	☐ doc#0	ire certainly more in the middle than we were previously because we are part of	BRICS	, along with Brazil, Russia, India and China.</s><s>We're in the BRICS club wit
2	☐ doc#0	a part of BRICS, along with Brazil, Russia, India and China.</s><s>We're in the	BRICS	club with Russia and China on the one hand and, of course, we still consider Eu
3	☐ doc#0	is and Cooperation, Naledi Pandor, to speak out against Russia's aggression in	BRICS	.</s><s>While, the Minister made quick mention of the situation, it is unfortunat
4	☐ doc#0	se of conquest of a democratic people and nation.</s><s>We are a member of	BRICS	[Brazil, Russia, India, China, and SA] and it should be used to condemn Russian
5	☐ doc#0	ise of conquest of a democratic people and nation.</s><s>We are a member of	BRICS	and it should be used to condemn Russian aggression. , the DA said the ANC g
6	☐ doc#0	condemn Russian aggression. , the DA said the ANC government must use its	BRICS	membership to condemn Russia's invasion of the Donbas region.</s><s>The pi
7	☐ doc#0	itial investment.</s><s>In the meanwhile, that country will remain a member of	BRICS	, and closely aligned with Russia, in a tangible way. # South Africa has given no
8	☐ doc#0	ny condemnation of the invasion that put them there. # Here's where the rest of	Brics	stands on Russia's invasion of Ukraine.</s><s>China: pro-Russia, with caveats
9	☐ doc#0	ter.</s><s>"There was a strong urge to abstain because SA is a member of the	Brics	[Brazil, Russia, India, China and South Africa] alliance.</s><s>It would have be
10	☐ doc#0	/s><s>It would have been awkward if it voted against Russia, which is its ally in	Brics	.</s><s>"South Africa ... cannot be punished for abstaining.</s><s>At the UN, t

Рис. 2. Конкорданс лексемы BRICS в корпусе ЮАР (Sketch Engine, Concordance)

Wordlist генерирует частотные списки различных типов: по части речи, начинающиеся/заканчивающиеся заданными морфемами, содержащие определенные символы, словоформы, леммы и другие атрибуты или комбинацию из трех вариантов, описанных выше. Частотный список корпуса ЮАР по части речи существительное представлен на рис. 3 для первых 20 единиц в порядке убывания частотности, что дает представление об общем содержании текстов корпуса.

Noun	Frequency ? ↓	Noun	Frequency ? ↓
1 russia	565 ...	11 sa	188 ...
2 africa	424 ...	12 us	184 ...
3 south	406 ...	13 world	175 ...
4 country	371 ...	14 india	162 ...
5 brics	361 ...	15 government	147 ...
6 china	270 ...	16 summit	143 ...
7 ukraine	209 ...	17 trade	132 ...
8 putin	205 ...	18 currency	123 ...
9 president	191 ...	19 ramaphosa	121 ...
10 war	190 ...	20 brazil	107 ...

Рис. 3. Частотный список существительных в корпусе ЮАР (Sketch Engine, Wordlist)

Keywords извлекает ключевые слова и термины (комбинации слов), типичные для исследуемого корпуса для корпуса/документа/текста или определяющие его содержание или тему, сопоставляя текст корпуса с референсным базовым корпусом соответствующего языка. Результат

разбивается на ключевые слова (элементы из одного слова) и термины (элементы из нескольких слов) и отображается со ссылками на предложения как в основном, так и в справочном корпусах. На рис. 4 представлены ключевые термины корпуса ЮАР, в качестве референсного корпуса Sketch Engine по умолчанию предложил корпус English Web (англоязычный корпус интернет-текстов общим объемом более 52 млрд слов¹).

Применение инструмента Sketch Engine к фрейм-анализу корпуса медиадискурса ЮАР играет важную роль, так как этот метод исследования фокусируется на изучении способов представления определенных тем, понятий или идей в дискурсе.

Term			Term		
1	brics summit	...	11	global south	...
2	brics country	...	12	brics currency	...
3	president cyril ramaphosa	...	13	anc government	...
4	russian president vladimir putin	...	14	fiat currency	...
5	international criminal court	...	15	arrest warrant	...
6	issue department	...	16	gold standard	...
7	russian invasion	...	17	rome statute	...
8	south africa	...	18	common currency	...
9	south african government	...	19	russian aggression	...
10	african government	...	20	new development bank	...

Рис. 4. Ключевые термины корпуса ЮАР (Sketch Engine, Keywords)

В этом контексте конкорданс KWIC может быть использован для идентификации и определения доминантных фреймов, их динамики, отражая изменения в общественном мнении или политической обстановке.

Следующим инструментом корпусного анализа, примененным к исследованию медиадискурсов стран БРИКС, стал инструмент **AntConc**², авторская программа разработчика Л. Антони. Интерфейс программы, доступной для большинства операционных систем, предлагает ряд инструментов, из которых в исследовании медиадискурсов стран БРИКС использовались следующие: Concordance, Cluster, N-Grams. Аналогично Sketch Engine, инструмент Concordance находит все контексты заданного слова или словосочетания в исследуемом корпусе, позволяя адаптировать размер контекста искомой единицы, а также осуществлять сортировку и фильтрацию полученных данных. Сопутствующая функция Concordance Plot представляет распределение необходимой единицы в тексте в виде штрих-кода, что позволяет визуально оценить ее частоту и местоположение (рис. 5).



Рис. 5. Конкорданс и распределение лексемы БРИКС в корпусе России (AntConc)

¹ enTenTen: Corpus of the English Web. URL: <https://www.sketchengine.eu/ententen-english-corpus/> (дата обращения: 11.09.2024).

² AntConc (Version 4.3.1). URL: <https://www.laurenceanthony.net/software> (дата обращения: 11.09.2024).

Cluster позволяет осуществлять отбор групп слов с определенным количеством элементов слева и справа от заданным лексемы, сортируя их по частотности и алфавиту (с начала или с конца кластера) (рис. 6).

Cluster	Rank	Freq	Range
брикс может стать альтернативой существующим	2	3	1
брикс повлияет на стабильность финансовой	2	3	1
брикс в июле года прошло	6	2	1
брикс в москве состоялась встреча	6	2	1
брикс вместе противодействуют наиболее опасным	6	2	1
брикс входят пять стран россия	6	2	1
брикс выпустит облигации в рублях	6	2	1
брикс выступили с решительным осуждением	6	2	1
брикс дал одобрение двум проектам	6	2	1

Рис. 6. Кластеры с начальной единицей БРИКС в корпусе России (AntConc)

Схожим функционалом располагает программа **LanA-Key**¹, разработанная компанией LanA Consulting. LanA-Key — это инструмент для извлечения ключевых слов/словосочетаний, который использует уникальный алгоритм интеллектуального извлечения именных словосочетаний, а также гибкие алгоритмы и параметры расчета статистики текста. В рамках анализа южноафриканского медиадискурса БРИКС данный инструмент использовался для экстракции ключевых фраз (n-grams), состоящих из одного или нескольких элементов — 1-граммов, 2-граммов, 3-граммов и 4-граммов (рис. 7). Источник текстов корпуса (южноафриканские ведущие СМИ) обуславливает высокую долю повторения элементов, содержащих в себе *South Africa/African*, а предметная область корпуса — повторения *BRICS* и названий стран-партнеров.

N-grams			
N-grClean	EM	Result	Index (General)
☒ Frequency ☐ ABC ☐ Reverse ABC			
Total 1-gr: 133554	Total 2-gr: 119122	Total 3-gr: 106067	Total 4-gr: 93233
Diff 1-gr: 12541	Diff 2-gr: 58770	Diff 3-gr: 82434	Diff 4-gr: 80597
the (9157) of (4402) and (4109) to (3944) in (3199) a (2523) is (1738) that (1458) for (1272) South (1185) as (1130) on (1004) with (937) Africa (915) it (846) has (778) this (752) be (752) are (689)	of the (1190) in the (785) South Africa (711) to the (423) and the (383) on the (298) for the (253) the BRICS (249) at the (233) with the (231) to be (219) It is (195) that the (193) by the (185) South African (185) as a (174) has been (172) as the (165) is a (161)	and South Africa (122) in South Africa (104) China and South (89) the South African (85) of the BRICS (59) one of the (57) South Africa is (52) new development Bank (50) the BRICS countries (49) South Africa has (46) of South Africa (41) South Africa and (41) the new development (40) there is a (38) in the world (38) the United states (34) in terms of (33) such as the (32) President Jacob Zuma (31)	China and South Africa (86) the new development Bank (35) the UN security Council (23) when it comes to (16) of the new development (16) of the South African (15) is one of the (15) of the BRICS countries (14) the South African government (13) one of the most (13) former President Jacob Zuma (13) at the University of (13) between South Africa and (12) at the end of (12) a member of the (12) I would like to (12) the democratic Republic of (11) the Africa regional Centre (11) Southern African development community (11)

Рис. 7. Ключевые слова и словосочетания корпуса ЮАП (LanA-Key)

В свою очередь, визуализация полученных данных осуществляется с применением инструмента **Voyant Tools**², который, помимо статистических функций, схожих с предыдущими инструментами (частотный анализ, ключевые слова, конкордансы и т.д.), позволяет исследователям создавать информативные визуализации, такие как облака тегов, диаграммы и карты, которые помогают наглядно представить данные. Voyant Tools имеет встроенные списки стоп-слов для большинства поддерживаемых языков, но пользователи также могут создавать собственные списки

¹ LanA-Key. URL: <http://www.lanaconsult.com/products.htm> (дата обращения: 11.09.2024).

² Voyant Tools. URL : <http://voyant-tools.org/> (дата обращения: 11.09.2024).

для более эффективного анализа. Среди функций визуализации выделяются Cirrus (облако наиболее частотных слов заданного корпуса), Links (сетевой график коллокаций), TermsBerry (визуализация высокочастотных терминов с их сочетаемостью) и др. Результаты обработки корпуса ЮАР с помощью указанных функций Voyant Tools представлены на рис. 8.

Опыт применения корпусных и компьютерных технологий в ходе исследования медиадискурсов стран БРИКС позволил представить их как эффективное средство объективной обработки больших объемов данных и, в связи с этим, как ценное дополнение к традиционным методам анализа. Перспективным представляется внедрение в исследование медиадискурсов стран БРИКС других корпусных компьютерных технологий.

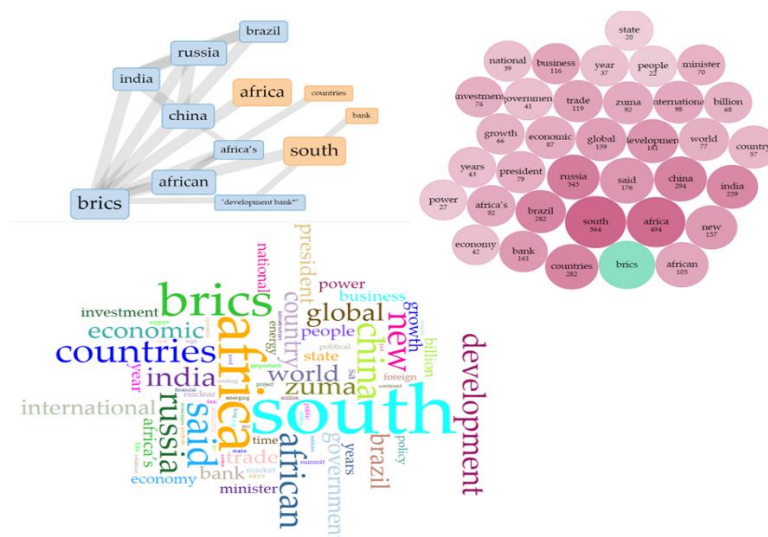


Рис. 8. Применение инструментов Cirrus, Links и TermsBerry к анализу корпуса ЮАР (Voyant Tools)

Одним из инструментов корпусного анализа, представляющих наибольший интерес, является UAM CorpusTool¹, разработанный компанией Wagsoft Linguistic Software, отличительной чертой которого является ориентация на работу с английским, китайским и португальским языком — основными языками при работе с дискурсами стран-партнеров БРИКС. Кроме того, привлечение нейронной сети «Концептоскоп»², разработанной учеными Кубанского государственного университета, призванной конструировать аналитическую модель исследовательского корпуса, в перспективе позволит визуализировать содержательную и тематическую структуру медиадискурса БРИКС.

СПИСОК ЛИТЕРАТУРЫ

1. Гнедаш, А. А. Опыт применения нейросети «Концептоскоп» для исследования процессов формирования гражданской идентичности в социальной сети Вконтакте / А. А. Гнедаш, В. В. Катермина // XII Столыпинские чтения. — Краснодар: Кубанский государственный университет, 2023. — С. 218-220.
2. Anthony, L. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit / L. Anthony // An Interactive Workshop on Language e-Learning. — 2004. — Pp. 7-13.
3. Thomas, J. Discovering English with the Sketch Engine / J. Thomas // CALL Design: Principles and Practice; Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands. — 2014. — Pp. 363-367.
4. O'Donnell, M. Demonstration of the UAM CorpusTool for text and image annotation / M. O'Donnell // Proceedings of the ACL-08: HLT Demo Session (Companion Volume). — 2008. — Pp. 13-16.

¹ UAM CorpusTool. URL: <http://www.corpustool.com/documentation.html> (дата обращения: 11.09.2024).

² Гнедаш А. А., Катермина В. В. Опыт применения нейросети «концептоскоп» для исследования процессов формирования гражданской идентичности в социальной сети Вконтакте // XII Столыпинские чтения. Краснодар: Кубанский государственный университет, 2023. С. 218–220.