ИЗМЕРЕНИЕ ЧАСТОТНОСТИ СИНТАКСИЧЕСКИХ МОЛЕКУЛ (НА МАТЕРИАЛЕ ГЕНЕРАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА) [1]

EVALUATING OF FREQUENCY OF SYNTACTIC MOLECULES (ON THE EVIDENCE FROM THE RUSSIAN GENERAL CORPUS)

Крылов С.А. *(krylov-58@mail.ru)*
Институт востоковедения РАН
Институт системного анализа РАН

Сделана попытка с помощью интегрированной информационной системы StarLing подсчитать частотность синтаксических молекул (минимальных знаменательных членов предложения, способных служить ответом на вопрос) на материале Генерального корпуса русского языка (созданного на основе Уппсальского корпуса).

1. Для грамматического и лексического анализа русского языка оказывается весьма полезным понятие синтаксической молекулы (СМ) [2]. СМ есть минимальная синтаксически автономная единица членения речи, то есть минимальный отрезок, способный функционировать в качестве отдельной (быть может и эллиптичной) реплики, отвечающей на какой-либо вопрос. СМ обычно содержит не более одного полнозначного знаменательного слова; при этом в её состав может входить одно или несколько служебных (или полуслужебных) слов.

2.0. Единица, близкая синтаксической молекуле, выделяется во многих фонетических работах под названием «фонетического слова» (ФС) [3] или «морфемного комплекса» [4]. Особенности предлагаемого подхода к ФС, предполагающего составление частотного словаря фонетических слов (Крылов 2008) – такие: (а) ФС рассматривается не только в синтагматическом, но и в парадигматическом аспекте; (б) ФС трактуется как двусторонняя (знаковая) единица [5]; (в) в центре внимания находится именно инвентарный [6] (словарный, лексикологический, лексикографический) аспект ФС [7].

2.1. В защиту принимаемого подхода можно привести следующие доводы, апеллирующие к аналогии между **слитно пишущимися** («синтетическими») речевыми отрезками и **раздельно пишущимися** («аналитическими») формами:

2.1.1. Если мы признаём «**синтетические**» латинские образования с постпозитивным *–que* двусторонними единицами, то можно ли отказывать в этом статусе «**аналитическим**» русским образованиям с препозитивным *и*?

2.1.2. Если мы признаём **синтетически** выражаемые категории падежа и числа (наряду с категориями рода и одушевлённости) категориями русских прилагательных, невзирая на то, что выбор граммем этих категорий обусловлен выбором соответствующих граммем существительного, контролирующего согласование данного прилагательного (и при этом готовы исключительно на основании такого контроля усматривать наличие соответствующих граммем в означаемом словоформ несклоняемых существительных типа *кофе* или *такси*, ср.: *чёрн-**ому** кофе*, *московск-**ого** такси*), то что нам мешает распространить этот подход на категории (и граммемы), выражаемые **аналитически** (напомним, что предлоги и послелоги, в соответствии со взглядами Ю. С. Маслова, признаются аналитическими выразителями категории **падежа**)? Соответственно, логично трактовать граммему ин-эссива как часть означаемого СМ ***в** письменн-**ом*** (признаваемую у СМ ***в** стол-**е***), граммему суб-латива – как часть означаемого СМ ***под** обеденн-**ым*** (признаваемую у СМ ***под** стол-**ом***) и т.п.

[1] Работа выполнена при частичной поддержке РГНФ (проект № 07-04-00161а.

[2] См. Балли 1955, с.317.

[3] См., напр., Зиндер 1979, с. 251; Суханова 1980, с. 90; Кодзасов и Кривнова 2001, с. 27-28, 304-306. В психолингвистическом аспекте важно проводимое Е. В. Ягуновой различение между более широким классом ФС «как оперативных единиц восприятия» и более узким классом ФС как «единиц перцептивного словаря» (Ягунова 2006, с. 400).

[4] Зализняк 1985: 39.

[5] Аргументация в пользу такого подхода приведена в Крылов 2007.

[6] Об «инвентарных» единиц языка в отличие от «конструктивных» см. Касевич 1988; Крылов 2006а.

[7] Такая трактовка созвучна идеям, развиваемым санкт-петербургской школой психолингвистически адекватного моделирования речевой деятельности. С другой стороны, проблему инвентаризации составных слов ставит и успешно решает практическая лексикография (Рогожникова 1991; Рогожникова 2003).

2.1.3. Если мы признаём двусторонними (знаковыми) единицами такие французские «**синтетические**» СМ, как *où* «, где (относит.), в котором (неодуш.)», *y* «в нём, у него (неодуш.)», *en* «от него (неодуш.)», то что нам мешает признавать двусторонними (знаковыми) единицами аналогичные им (а при известных условиях даже эквивалентные им) русские «**аналитические**» СМ: *в котором*, *у которого*, *в нём*, *у него*, *от него*, *из него* и т.п.?

2.1.4. Если мы признаём двусторонними (знаковыми) единицами такие русские «**синтетические**» образования (1а) *там*; (1б) *туда*; (1в) *оттуда*; (2а) где (относит.); (2б) *куда* (относит.); (2в) *откуда* (относит.); и т. п., то что нам мешает признать двусторонними (знаковыми) единицами их «**синтетические**» аналоги (а нередко и контекстные эквиваленты) в русском языке (1а) *в нём*, *в ней*, *на нём*, *на ней*, *у него*, *у неё*, *у них*; (1б) *в него*, *в неё*, *в них*; *на него*, *на неё*, *на них*; *к нему*, *к ней*, *к ним*; (1в) *из него*, *из неё*, *из них*; *с него*, *с неё*, *с них*; *от него*, *от неё*, *от них*; (2а) *в котором*, *в которой*, *в которых*; *на котором*, *на которой*, *на которых*; *у которого*, *у которой*, *у которых*; (2б) *в который*, *в которую*, *в которые*; *на который*, *на которую*, *на которые*; *к которому*, *к которой*, *к которым*; (2в) *из которого*, *из которой*, *с которого*, *с которой*, *с которых*; *от которого*, *от которой*, *от которых*; и т. п.)?

2.1.5. Если мы признаём двусторонней (знаковой) единицей «**синтетически**» образованную форму с деепричастным значением (ср. *читая*), то что нам мешает признать двусторонней (знаковой) единицей аналогичную ей (и нередко контекстно эквивалентную ей) «**аналитически**» образованную СМ с сочинительным значением (ср. *и читает*)?

2.2. Если мы признаём двусторонней (знаковой) единицей «**аналитически**» образованную форму с комитативным значением, оформляющую неоднородный член (напр., во фразе *Петя* ***с Ваней*** *гуляют*), то что нам мешает признать двусторонней (знаковой) единицей эквивалентную ей «**аналитически**» же образованную форму с координативным значением, оформляющую однородный член (напр., во фразе *Петя* ***и Ваня*** *гуляют*)?

3.0. СМ образуют иерархию из трёх рангов [8]: макротакты, мезотакты и микротакты.

3.1. Макротакт есть морфемный комплекс между двумя местами потенциальных пауз (в отличие от более крупной единицы - фонетической синтагмы[9], границы которой отмечены реальными паузами).

3.2. Мезотакт есть морфемный комплекс, включающий не более одного «полноударного» ФС. Мезотакт может включать в себя один или несколько **«клитикоидов» [10]** (то есть **«слабоударяемых»** ФС и «относительных клитик») – постпозитивных (**«энклитикоидов»**) или препозитивных (**«проклитикоидов»**).

3.3. Микротакт есть морфемный комплекс, содержащий ровно 1 автономный (характеризуемый единством главного словесного ударения) словесный сегмент[11]. Микротакты бывают простыми и составными. Составные микротакты включают, помимо автономного сегмента, также одну или несколько **клитик** – единиц, не несущих самостоятельного словесного ударения. Клитики подразделяются на **энклитики** (постпозитивные) и **проклитики** (препозитивные).

4.0. Инвентарь ментальных СМ (то есть синтаксических молекул, хранимых в ментальном лексиконе носителя языка) выявляется путём измерения их встречаемости в крупном корпусе текстов [12] и создания частотного инвентаря реальных СМ [13].

4.1. Эта задача может решаться по-разному [14]. Источником данных был корпус текстов, представленных в орфографической записи - Генеральный корпус русского языка (ГКРЯ), созданный на основе «Уппсальского корпуса» русского языка (УпКРЯ), составленного под руководством Л. Лённгрена (http://www.slaviska.uu.se/ryska/index.html). В 1995 гг. автором настоящей работы под руководством С. А. Старостина (1953-2005) [15] материалы УпКРЯ были преобразованы в формат текстовой базы данных, получившей название ГКРЯ [16].

5.0. В 2005-2008 гг. ГКРЯ был снабжён «грубой» разметкой тактовой делимитации. Она устроена так.

5.1. Пробелы письменного текста бывают **паузальные** (соответствующие границам макротактов в устной

[8] Эти понятия введены в Крылов 2006b, Крылов 2006c.

[9] О фонетических синтагмах см. Кодзасов и Кривнова 2001, с. 2728, 304-306.

[10] Термины из Крылов 2006b, Крылов 2006c.

[11] Иными словами, микротакт соответствует «тактовой группе» (Чурганова 1973, с. 22-23, 27, 31-32), «акцентному слову» (Маслов 1975, с. 91-93), «акцентной группе» (Зиндер 1979, с. 251) «речевому такту» (Кодзасов и Кривнова 2001, с. 306), «ритмической группе» (Зиндер 1979, с. 250, 262; Кодзасов и Кривнова 2001, с. 310).

[12] Т.о., разрабатывается «структурно-вероятностная модель» языка (Шайкевич 1990: 231).

[13] См. Крылов и Ягунова 2006.

[14] В современной корпусной лингвистике иногда не ограничиваются членением текста на графические слова (от пробела до пробела), а выделяют также более сложные единицы – т.н. «составные слова». См., напр., Венцов и др. 2004a; Венцов и др. 2004b; Копотев 2004; Мустайоки и Копотев 2004; Ягунова 2006.

[15] Взгляды С. А. Старостина на задачи корпусной филологии отразились в публикации Перцов и Старостин 1995; см. также проект ПРИОР «Портал “Русский язык и литература”» (http://prior.russia-gateway.ru/content/view/1023/150/).

[16] См. Крылов и Старостин 2005.

речи) и **беспаузальные** (для транскрибирования которых использован создан набор из 6 искусственных делимитаторов [17]:

{ после проклитик;

} перед энклитикой;

< после проклитикоида;

> перед энклитикоидом;

<> между частями мезотакта с «неустойчивым» центром (то есть сочетания, допускающего двоякую акцентуацию: либо как «клитикоид + полноударное», либо как «полноударное + клитикоид»);

+ между мезотактами, образующими один макротакт.

6. Частотность СМ в русском языке.

В результате разметки ГКРЯ оказалось возможным извлечь из него сведения о частотах СМ.

Сосредоточим внимание на одном из классов СМ – а именно, на СМ, начинающихся с проклитики. Рассмотрим небольшую «верхушку» из частотного словаря (ЧС) таких СМ.

В таблице столбец (А) указывает на инвентаризуемую СМ (макротакт), (Б) - на её относительную частотность по числу текстов (%), (В) - на её абсолютную частотность по числу текстов, (Г) - на её ранг в ЧС, упорядоченном по числу текстов (этот параметр в таблице является ключевым), (Д) - на её относительную частотность по числу вхождений при измерении общего числа вхождений СМ в корпус (в числе вхождений данной единицы на 10 тыс. [18]), (Е) - на её абсолютную частотность по числу вхождений (этот параметр в таблице является побочным), (Ж) - на её ранг в ЧС, упорядоченном по числу вхождений. Под ключевым параметром понимается та из числовых характеристик СМ, которая лежит в основе упорядочения строк в таблице (в нашем случае в качестве ключевого был выбран параметр Г).

Для наглядности ниже дана лишь частотная «верхушка» одного из полученных словарей [19].

Частотность мезотактов с проклитиками в ЧС макротактов.

А	Б	В	Г	Д	Е	Ж
о{том	35.98	204	61	44.75	319	86
у{нас	30.51	173	87	44.61	318	87
из{них	25.93	147	121	28.90	206	161
об{этом	25.75	146	124	28.90	206	163
не{}было	24.34	138	138	48.26	344	76
в{нем	22.22	126	165	28.20	201	170
и{все	21.87	124	168	29.60	211	155
и{это	20.99	119	190	22.31	159	249
у{него	20.28	115	200	43.35	309	90
а{потом	19.93	113	207	30.72	219	145
и{другие	19.75	112	219	18.80	134	316
с{ним	19.58	111	220	26.37	188	186
к{нему	19.05	108	231	24.97	178	210
в{ней	18.87	107	235	20.06	143	286
и{его	18.87	107	236	17.11	122	378
в{котором	18.17	103	247	17.82	127	352
у{них	17.81	101	254	21.60	154	263
в{частности	17.46	99	264	22.87	163	241
и{что	16.93	96	286	19.36	138	302
к{сожалению	16.75	95	301	16.41	117	404
на{него	16.05	91	312	22.59	161	243
у{нее	15.87	90	319	29.60	211	157
у{меня	15.87	90	320	25.11	179	208
и{как	15.52	88	332	16.27	116	408
до{сих{пор	15.52	88	336	15.43	110	451
к{ней	15.34	87	340	19.64	140	292
и{других	15.34	87	344	15.43	110	453
не{может	15.34	87	346	14.73	105	480
в{них	15.17	86	349	14.87	106	468
в{целом	14.81	84	359	15.85	113	428

[17] О просодических швах разной глубины см. Кривнова 2007: 60-69.

[18] Ср. понятие «ipm» (“instances per million words”) в Sharoff 2002 (http://www.artint.ru/projects/frqlist.asp).

[19] Развёрнутые версии этой и других таблиц (проекция срезов ЧС 4096 наиболее частых макротактов и 4096 самых частых микротактов) выложены в Сети (http://www.yazykoznanie.narod.ru/PHOWO08.html).

на{себя	14.81	84	360	15.57	111	443
на{них	14.64	83	368	14.59	104	486
к{тому}же	14.64	83	369	13.19	94	550
а{это	14.46	82	375	13.61	97	530
и{так	14.11	80	390	15.71	112	437
в{мире	14.11	80	393	14.17	101	501
а{что	13.76	78	405	14.45	103	490
в{*Москве	13.58	77	413	13.61	97	531
и{вдруг	13.23	75	427	17.40	124	370
в{стране	13.23	75	428	16.69	119	393
в{год	13.23	75	430	15.43	110	446
в{которой	13.23	75	440	12.06	86	621
к{ним	12.87	73	456	12.49	89	591
в{сторону	12.70	72	465	13.75	98	521
и{снова	12.52	71	471	14.73	105	479
и{тогда	12.35	70	485	14.03	100	506
и{они	12.35	70	486	13.75	98	522
во{всех	12.35	70	489	10.80	77	726
и{т#+д#	12.17	69	491	15.43	110	454
а{он	12.17	69	493	14.59	104	483
в{жизни	12.17	69	496	12.91	92	564
как{правило	12.17	69	498	12.20	87	610
не{будет	12.17	69	501	11.36	81	678
не{мог	11.82	67	524	15.15	108	462
и{теперь	11.82	67	526	12.49	89	590
в{которых	11.82	67	531	10.80	77	724
и{она	11.64	66	533	17.25	123	373
а{затем	11.46	65	564	10.94	78	710
от{него	11.29	64	570	13.19	94	554
к{себе	11.29	64	574	11.50	82	663
в{результате	11.29	64	576	10.52	75	752
с{ними	11.11	63	585	11.78	84	647
к{примеру	11.11	63	590	11.08	79	697
во{всем	11.11	63	592	10.66	76	737
а{я	10.93	62	597	14.73	105	475
в{себе	10.93	62	605	11.36	81	673
в{первую+очередь	10.93	62	612	9.96	71	821
и{потому	10.76	61	620	10.52	75	756
а{теперь	10.76	61	621	10.38	74	769
в{основном	10.76	61	625	9.68	69	857
и{тут	10.58	60	631	11.50	82	662
и{их	10.58	60	635	9.82	70	842
и{когда	10.41	59	651	11.64	83	652
с{ней	10.41	59	652	11.50	82	668
в{чем	10.41	59	657	10.24	73	787
для{него	10.41	59	661	9.96	71	825
на{нее	10.23	58	671	12.63	90	586
и{ее	10.23	58	679	10.24	73	788
а{когда	10.05	57	699	10.52	75	750
и{сейчас	10.05	57	704	8.70	62	1001
и{я	9.88	56	706	23.15	165	233
о{чем	9.70	55	741	10.66	76	745
в{нашей<стране	9.52	54	762	9.40	67	896
до{конца	9.52	54	765	9.26	66	914
по{существу	9.52	54	768	8.42	60	1049
а{тут	9.35	53	779	9.12	65	932
не{так	9.35	53	781	8.70	62	1006
для{себя	9.17	52	807	8.84	63	978
в{прошлом<>году	8.99	51	829	9.54	68	880
от{нее	8.99	51	833	8.84	63	987
не{знаю	8.82	50	850	9.96	71	832
а{она	8.82	50	852	9.40	67	894
не{раз	8.82	50	860	8.56	61	1026
на{месте	8.82	50	863	8.28	59	1070
тем<не{менее	8.82	50	865	8.28	59	1081
во{многом	8.82	50	866	8.14	58	1087
не{могут	8.82	50	870	8.00	57	1119
и{уже	8.82	50	872	7.86	56	1149
от{них	8.82	50	874	7.58	54	1230
в{последнее<>время	8.64	49	902	7.86	56	1142
из{которых	8.64	49	905	7.72	55	1177
в{*C*C*C*P	8.47	48	913	10.66	76	736
в{общем	8.47	48	917	9.26	66	911

в{руках	8.47	48	923	8.28	59	1060
а{значит	8.47	48	931	7.44	53	1239
для{них	8.47	48	932	7.29	52	1288
а{может	8.29	47	941	8.84	63	974
а{ты	7.94	45	983	10.94	78	711
в{конце<>концов	7.94	45	994	8.28	59	1059
и{все}же	7.94	45	995	8.28	59	1062
и{мы	7.94	45	996	8.28	59	1063
с{собой	7.94	45	997	8.28	59	1077
и{вообще	7.94	45	1001	7.72	55	1174
и{сам	7.94	45	1002	7.72	55	1175
для{всех	7.94	45	1012	7.15	51	1324
и{наконец	7.94	45	1013	7.15	51	1326
не{надо	7.76	44	1036	8.00	57	1120
на{землю	7.76	44	1039	7.72	55	1183
в{одном	7.76	44	1050	7.15	51	1319
в{самом	7.76	44	1052	7.01	50	1351
в{то}же<>время	7.76	44	1062	6.59	47	1464
не{всегда	7.76	44	1063	6.45	46	1527
в{работе	7.58	43	1077	7.72	55	1167
о{нем	7.58	43	1080	7.44	53	1261
и{тут}же	7.58	43	1082	7.29	52	1290
на{все	7.58	43	1086	6.87	49	1406
в{свое<время	7.58	43	1088	6.59	47	1463
в{*С*Ш*А	7.41	42	1103	8.00	57	1110
а{как	7.41	42	1113	7.15	51	1316
во{все	7.41	42	1126	6.45	46	1505
на{котором	7.41	42	1127	6.45	46	1524
в{таких	7.41	42	1128	6.31	45	1549
на{нем	7.41	42	1132	6.17	44	1600
о{них	7.41	42	1133	6.17	44	1606
за{ним	7.23	41	1141	9.12	65	935
и{все-таки	7.23	41	1154	7.01	50	1356
на{другой	7.23	41	1157	6.87	49	1407
на{этом	7.23	41	1158	6.73	48	1442
во{всяком<>случае	7.23	41	1162	6.45	46	1506
на{меня	7.05	40	1175	8.98	64	960
у{вас	7.05	40	1180	8.00	57	1133
на{улице	7.05	40	1185	7.58	54	1221
со{мной	7.05	40	1188	7.44	53	1269
о{себе	7.05	40	1191	7.29	52	1293
в{глаза	7.05	40	1195	7.15	51	1318
к{чему	7.05	40	1214	6.03	43	1654
в{городе	6.88	39	1219	8.98	64	954
для{меня	6.88	39	1223	8.28	59	1061
и{т#п#	6.88	39	1224	8.28	59	1064
и{стал	6.88	39	1229	7.72	55	1176
в{настоящее<>время	6.88	39	1234	7.29	52	1280
не{знал	6.88	39	1237	7.01	50	1365
на{берегу	6.88	39	1239	6.87	49	1405
и{там	6.88	39	1244	6.59	47	1469
а{сейчас	6.88	39	1249	6.45	46	1500
в{него	6.88	39	1252	6.31	45	1548
ко{мне	6.70	38	1270	8.00	57	1118
на{земле	6.70	38	1279	7.01	50	1363
на{работу	6.70	38	1280	7.01	50	1364
в{день	6.70	38	1286	6.73	48	1427
для{нас	6.70	38	1289	6.59	47	1467
на{всех	6.70	38	1301	6.03	43	1658
не{все	6.70	38	1302	6.03	43	1662
не{хватает	6.70	38	1306	5.89	42	1710
в{школе	6.53	37	1315	11.08	79	694
с{тех<>пор	6.53	37	1335	6.59	47	1491
с{одной>стороны	6.53	37	1341	6.31	45	1567
и{в{то}же<>время	6.53	37	1344	6.17	44	1593
из{нас	6.53	37	1345	6.17	44	1595
со{всеми	6.53	37	1346	6.17	44	1627
на{этот<раз	6.35	36	1386	6.59	47	1475
а{может<>быть	6.35	36	1400	5.89	42	1682
в{себя	6.35	36	1405	5.75	41	1742
и{есть	6.35	36	1407	5.75	41	1759
за{собой	6.35	36	1414	5.61	40	1816
в{нее	6.35	36	1416	5.47	39	1865

и{поэтому	6.35	36	1418	5.47	39	1880
из{этих	6.35	36	1419	5.47	39	1881
а{потому	6.17	35	1437	7.58	54	1199
в{воздухе	6.17	35	1446	6.59	47	1462
в{памяти	6.17	35	1454	6.03	43	1637
и{тоже	6.17	35	1458	5.89	42	1697
в{разных	6.17	35	1462	5.75	41	1741
не{может<>быть	6.17	35	1463	5.75	41	1770
в{этом<случае	6.17	35	1477	5.33	38	1929
о{котором	6.17	35	1479	5.33	38	1959
в{свою<очередь	6.17	35	1481	5.19	37	2009
с{таким	6.17	35	1484	5.05	36	2130
у{тебя	6.00	34	1494	7.01	50	1380
к{нам	6.00	34	1498	6.45	46	1516
не{могу	6.00	34	1503	6.17	44	1603
и{сразу	6.00	34	1519	5.75	41	1760
с{которой	6.00	34	1524	5.47	39	1902
в{данном<>случае	6.00	34	1525	5.33	38	1925
по{всем	6.00	34	1529	5.33	38	1965
на{всю	6.00	34	1536	5.05	36	2096
со{временем	6.00	34	1540	5.05	36	2134
то<и{дело	5.82	33	1561	5.89	42	1728
на{которой	5.82	33	1568	5.61	40	1832
в{самом<>деле	5.82	33	1574	5.47	39	1866
друг<от{друга	5.82	33	1576	5.47	39	1876
не{случайно	5.82	33	1577	5.47	39	1888
в{последние<>годы	5.82	33	1579	5.33	38	1927
на{ней	5.82	33	1581	5.33	38	1953
в{два	5.82	33	1587	5.19	37	2005
у{всех	5.82	33	1592	5.19	37	2053
а{они	5.82	33	1594	5.05	36	2062
на{что	5.82	33	1597	5.05	36	2097
с{другой<>стороны	5.82	33	1604	4.91	35	2212
в{доме	5.64	32	1614	7.86	56	1141
и{опять	5.64	32	1617	7.44	53	1250
на{свете	5.64	32	1623	6.45	46	1525
не{могла	5.64	32	1627	6.17	44	1602
на{самом<>деле	5.64	32	1633	5.89	42	1708
в{лицо	5.64	32	1637	5.61	40	1809
в{сущности	5.64	32	1640	5.47	39	1867
в{одну	5.64	32	1653	5.19	37	2007
ни{разу	5.64	32	1656	5.19	37	2027
в{дальнейшем	5.64	32	1662	5.05	36	2063
в{одной	5.64	32	1663	5.05	36	2065
в{руки	5.64	32	1664	5.05	36	2066
и{потом	5.64	32	1673	4.91	35	2179
не{имеет	5.64	32	1675	4.91	35	2191
и{больше	5.64	32	1680	4.77	34	2250
а{где	5.64	32	1687	4.63	33	2304
не{могли	5.64	32	1688	4.63	33	2342
в{комнате	5.47	31	1711	5.89	42	1686
друг<с{другом	5.47	31	1712	5.89	42	1694
в{истории	5.47	31	1719	5.75	41	1739
и{сказал	5.47	31	1724	5.61	40	1821
а{мы	5.47	31	1730	5.33	38	1922
в{другую	5.47	31	1740	5.05	36	2064
по{крайней<>мере	5.47	31	1745	5.05	36	2115
и{дальше	5.47	31	1763	4.63	33	2326
с{трудом	5.47	31	1767	4.63	33	2363
на{улицу	5.29	30	1775	6.87	49	1408
а{вы	5.29	30	1786	6.03	43	1635
в{природе	5.29	30	1793	5.75	41	1740
на{столе	5.29	30	1801	5.61	40	1833
а{главное	5.29	30	1813	5.05	36	2061
о{своем	5.29	30	1814	5.05	36	2101
в{среднем	5.29	30	1820	4.91	35	2161
а{сам	5.29	30	1823	4.77	34	2231
на{одном	5.29	30	1826	4.77	34	2253
и{тем<не{менее	5.29	30	1831	4.63	33	2327
в{голову	5.29	30	1835	4.49	32	2394
в{различных	5.29	30	1836	4.49	32	2396
в{системе	5.29	30	1837	4.49	32	2397
время<от{времени	5.11	29	1868	5.47	39	1870

в{*Москву	5.11	29	1871	5.33	38	1924
в{один	5.11	29	1893	4.91	35	2160
с{детства	5.11	29	1897	4.91	35	2211
а{все	5.11	29	1917	4.49	32	2387
в{этом<году	5.11	29	1920	4.49	32	2399
в{другом	5.11	29	1932	4.35	31	2489
а{иногда	5.11	29	1937	4.21	30	2564
в{каждой	5.11	29	1938	4.21	30	2570
в{этом	5.11	29	1939	4.21	30	2573
к{концу	5.11	29	1943	4.21	30	2607
не{столько	5.11	29	1945	4.21	30	2626

Заключение

Приведённые данные являются сугубо предварительными, так как процесс разметки ГКРЯ пока продолжается. Многие из выделенных синтаксических молекул нуждаются в более тщательной интерпретации. Но можно надеяться, что и эта предварительная стадия разметки способна привести к получению ценной информации о статистике употребления СМ в русском языке.

Список литературы

1. Аванесов Р. И. Русское литературное произношение. М.: Просвещение, 1968.- 288 с.
2. Аванесов Р. И. Русская литературная и диалектная фонетика. М.: Просвещение, 1974.- 287 с.
3. Балли Ш. Общая лингвистика и вопросы французского языка. М.: ИЛ, 1955.- 416 с.
4. Венцов А. В., Касевич В. Б., Ягунова Е. В. Идиома, слово, фонетическое слово // Язык и речь: проблемы и решения. Сб. научных трудов к юбилею проф. Л. В. Златоустовой. М.: Изд-во МГУ, 2004a. С. 357-363.
5. Венцов А. В., Грудева Е. В., Касевич В. Б., Ягунова Е. В. Об идиомах в национальном корпусе русского литературного языка // Корпусная лингвистика-2004. Тезисы международной конференции. 12-14 октября 2004 г., СПб.: Изд. СпбГУ, 2004b. С. 1718.
6. Высотский С. С. Звук речи в контексте // Диалектологические исследования по русскому языку. М.: Наука, 1977. С. 2438.
7. Зализняк А. А. От праславянской акцентуации к русской. М.: Наука, 1985.- 428 с.
8. Зиндер Л. Р. Общая фонетика. Изд. 2-е. М.: Высшая школа, 1979.- 312 с.
9. Касаткин Л. Л. Фонетика современного русского литературного языка. М.: Изд-во МГУ, 2003.- 223 с.
10. Касевич В. Б. Семантика. Синтаксис. Морфология. М.: Восточная литература, 1988.
11. Кодзасов С. В., Кривнова О. Ф. Общая фонетика. М.: РГГУ, 2001.- 592 с.
12. Копотев М. Несмотря на, потому что, или многокомпонентные единицы в аннотированном корпусе русских текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2004» («Верхневолжский», 2-7 июня 2004 г.), М., Наука, 2004.
13. Кривнова О. Ф.Ритмизация и интонационное членение текста в «процессе речи-мысли». ДД. М., 2007.
14. Крылов С. А.Об инвентарных и конструктивных единицах языка // Язык и речевая деятельность. 2003. Вып. 6. СПб.: Филол. ф-т СПбГУ, 2006a. С. 926.
15. Крылов С. А.Фонетическое слово и его корреляты в русском письменном тексте (с точки зрения корпусной лингвистики) // Корпусная лингвистика-2006. Труды международной конференции. 10-14 октября 2006 г., СПб.: Изд. СпбГУ РХГА, 2006b. С. 190200.
16. Крылов С. А. Фонетическое слово и его корреляты в русском письменном тексте (с точки зрения корпусной лингвистики) // Пятая выездная школа-семинар «Порождение и восприятие речи». Череповец: ЧГУ – ПЛО, 2006c. С. 6696.
17. Крылов С. А. Фонетическое слово в семиотическом аспекте // Фонетика сегодня. Материалы докладов и сообщений V международной научной конференции. 8-10 октября. М.: ИРЯ РАН, 2007. С. 210212.
18. Крылов С. А., Старостин С. А. Металингвистическая разметка текстовых баз данных в системе STAR-LING и современные задачи корпусной лингвистики // Прикладная лингвистика в поиске новых путей. Тезисы Международной научной конференции MegaLing'2005. Крым, Украина. 27 июня - 2 июля 2005 г. Симферополь: ТЭИ, 2005.
19. Крылов С. А. , Ягунова Е. В. 2006. Квантитативный подход к выделению инвентарных единиц языка // Материалы 2-й международной конференции по когнитивной науке. СПб., 9-13 июня 2006 года. СПб., 2006. С. 329-330.
20. Маслов Ю. С. Введение в языкознание. М.: Высшая школа, 1975.

21. Мустайоки А., Копотев М. К вопросу о статусе эквивалентов слова типа потому что, в зависимости от, к сожалению // Вопросы языкознания, 2004, № 3.

22. Перцов Н.В., Старостин С. А. О лексикографической справочной информационной системе ЛЕКСИС по русскому языку // Труды Международного семинара «Диалог'95» по компьютерной лингвистике и ее приложениям = «Dialogue'95. Computational linguistics and its applications» international workshop, Казань, 31 мая - 4 июня 1995 г. - Казань, 1995. - С. 247.

23. Рогожникова Р. П. Словарь эквивалентов слова: наречные, служебные, модальные единства. - М.: Русский язык, 1991. - 255 с.

24. Рогожникова Р.П. Толковый словарь словосочетаний, эквивалентных слову. М., 2003.

25. Русская грамматика. Т. 1. М.: Наука, 1980.- 783 с.

26. Суханова М. С. Основные сведения об ударении // Русская грамматика 1980. С. 9095.

27. Чурганова В. Г. Очерк русской морфонологии. М.: Наука, 1973.- 239 с.

28. Шайкевич А. Я. Количественные методы в языкознании // Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. С. 231.

29. Ягунова Е. В. Неоднословные целостности в словаре и в корпусе // Корпусная лингвистика-2006. Труды международной конференции. 10-14 октября 2006 г., СПб.: Изд. СпбГУ РХГА, 2006. С. 395412.

30. Sharoff, Serge, Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. // Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain, 2002.