

**Измерение частотности
синтаксических молекул (на
материале Генерального
корпуса русского языка)**

***С. А. Крылов
(krylov-58@mail.ru)***

ПОНЯТИЕ СИНТАКСИЧЕСКОЙ МОЛЕКУЛЫ

- 1. Для грамматического и лексического анализа русского языка оказывается весьма полезным понятие синтаксической молекулы (СМ).
- СМ есть минимальная синтаксически автономная единица членения речи, то есть минимальный отрезок, способный функционировать в качестве отдельной (быть может и эллиптической) 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.

«синтаксическая молекула» и «фонетическое слово»

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

3.0. Три ранга СМ: макротакты, мезотакты и микротакты

- **3.1. Макротакт** – морфемный комплекс между двумя местами потенциальных пауз (в отличие от более крупной единицы - фонетической синтагмы, границы которой отмечены реальными паузами).
- **3.2. Мезотакт** – морфемный комплекс, включающий не более одного «полноударного» ФС. Мезотакт может включать в себя один или несколько «**клитикоидов**» (то есть «**слабоударяемых**» ФС и «**относительных клитик**») – **постпозитивных («энклитикоидов»)** или **препозитивных («проклитикоидов»)**.
- **3.3. Микротакт** – морфемный комплекс, содержащий ровно 1 автономный (характеризуемый единством главного словесного ударения) словесный сегмент. 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).

способы выявления инвентаря «ментальных СМ»

- **4.0.** Инвентарь ментальных СМ выявляется путём измерения их встречаемости в крупном корпусе текстов и создания частотного инвентаря реальных СМ .
- **4.1.** Эта задача может решаться по-разному. Источником данных был корпус текстов, представленных в орфографической записи -- Генеральный корпус русского языка (ГКРЯ), созданный на основе «Упсальского 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) материалы УпКРЯ были преобразованы в формат текстовой базы данных, получившей название ГКРЯ.

принципы «грубой» разметкой тактовой делимитации

- **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 речи) и **беспаузальные** (для транскрибирования которых использован создан набор из 6 искусственных делимитаторов:
 - { после проклитик;
 - } перед энклитикой;
 - < после проклитикоида;
 - > перед энклитикоидом;
 - <> между частями мезотакта с «неустойчивым» центром (то есть сочетания, допускающего двойную акцентуацию: либо как «клитикоид + полноударное», либо как «полноударное + клитикоид»);
 - + между мезотактами, образующими один макротакт.

таблица «Частотность мезотактов с проклитиками в ЧС макротактов»

- **6.0.** В таблице столбец (А) указывает на инвентаризируемую 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., (Е) - на её абсолютную частотность по числу вхождений (этот параметр в таблице является побочным), (Ж) - на её ранг в ЧС, упорядоченном по числу вхождений.

- 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.
- Для наглядности ниже дана лишь частотная «верхушка» одного из полученных словарей

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	550 263
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
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

						673
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	896 821
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 ¹¹⁷⁷
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 {СССР	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

9551 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 {США	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