Роберт Колар, Петр Плехач
Институт чешской литературы АН ЧР
(Чехия, Прага)
kolar@ucl.cas.cz, plechac@ucl.cas.cz

ЧАСТОТНОСТЬ ЧАСТЕЙ РЕЧИ В ЧЕШСКОЙ ПОЭЗИИ*

В статье представлены первые результаты статистической обработки Корпуса чешского стиха. Корпус чешского стиха является лемматизированным, фонетически, морфологически, метрически и строфически аннотированным корпусом чешской поэзии периода XIX — начала XX вв. и содержит 1700 сборников, около 15 миллионов слов, 2,5 миллиона стихов. В первой части статьи сравнивается частотность частей речи в Корпусе чешского стиха и в доступных корпусах чешского языка (включая подкорпусы художественной литературы, публицистических и специальных текстов, разговорного языка). Затем анализируется зависимость частотности частей речи от длины стиха (частотность личных форм глагола, именной и глагольной частей предложения), поэтической школы (чешская поэзия 50–60-х и 70–80-х гг. XIX в.), автора (на основе противопоставления творчества Сватоплука Чеха и Виктора Дыка) и рода литературы (на примере лирических и эпических произведений Адольфа Гейдука). По причине ограниченности проанализированного материала результаты исследования не являются окончательными, однако последующее изучение более обширного материала позволит использовать различия в значениях наблюдаемых параметров в качестве стилеметрических показателей как в истории литературы, так и при атрибуци в текстологии.

Ключевые слова: стихосложение, стилеметрия, корпусная лингвистика, части речи, поэтика

Введение

Целью настоящей статьи является представление первых результатов статистической обработки Корпуса чешского стиха (Korpus českého verše, далее КЧС). КЧС основывается на материалах Чешской электронной библиотеки (Česká elektronická knihovna, далее ЧЭБ), которая представляет собой открытую полнотекстовую базу

* Данное исследование проводилось в 2012 г. при поддержке Агентства грантов Чешской Республики, грант №P406/11/1825, а также при поддержке, направленной на долговременное концептуальное развитие исследовательского учреждения 68378068.

данных, содержащую около 1700 сборников[1], т. е. более 13 миллионов слов и примерно 2,5 миллиона стихов (см. www.ceska-poezie.cz). КЧС был автоматически аннотирован (с фонетической, морфологической, метрической и строфической точек зрения, ср.: Ibrahim — Plecháč 2011). В предлагаемом ниже тексте мы ограничиваемся анализом частотности употребления частей речи[2].

Исходным пунктом данному исследованию послужили работы Л. Пшчоловской [Pszczołowska 1965], С. Мазачовой [Mazáčová 1973], М. Червенки и К. Сгалловой [Červenka — Sgallová 1984], М. Л. Гаспарова и Т. В. Скулачевой [Гаспаров — Скулачева 2004].

В Чехии частотностью употребления частей речи в художественной литературе занималась М. Тешитэлова [Těšitelová 1968, 1974, 2000]. Тешитэлова в своих работах исходила, прежде всего, из так называемого частотного словаря чешского языка [далее ЧСЧЯ, ср.: Jelínek 1961], основывающегося на 75 произведениях (восьми подкорпусах: художественная проза, поэзия, молодежная литература, драма, специальная литература, публицистика, научная литература, разговорная речь), изданных преимущественно в 30–40-е годы XX века; поэзия в нем представлена только десятью сборниками. Наряду с данными из ЧСЧЯ мы приводим данные из корпуса нехудожественных стилей (далее НС), который включает подкорпусы публицистического, научного и официально-делового стилей; письменные и разговорные тексты в нем представлены в соотношении 75 % : 25 % [Těšitelová 1985]. Однако данные из ЧЭБ мы сравнивали, прежде всего, с данными публикации «Статистика чешского языка» [Bartoň 2009] и, отчасти, с данными из публикации «Морфология разговорного чешского языка: частотный анализ» [Šonková 2008]. Данные публикации «Статистика чешского языка» основываются на Чешском национальном корпусе SYN2005 (см. Český národní korpus), охватывающем современный письменный чешский язык и состоящем из трех подкорпусов (художественная проза, специальные и публицистические тексты); эти данные мы пополнили статистикой по 90 стихотворным текстам, в том числе переводным, изданным в 90-е годы, которые также содержатся в SYN2005. Данные «Морфологии разговорного чешского языка» основываются на Пражском разговорном корпусе (Pražský mluvený korpus, далее ПРК), включающем в себя речевые записи 504 говорящих из Праги и близлежащей области, записанные на пленку в период с 1988 по 1996 гг.

[1] В подавляющем большинстве случаев одному наименованию соответствует один сборник стихов, однако некоторые наименования дублируются (например, первое издание сборника и его издание из сочинений данного автора). Учитывая это, мы отстранили из корпуса некоторые наименования для нужд данного исследования, в связи с чем количество проанализированных наименований составило лишь немногим более полутора тысяч.

[2] Лемматизацию и морфологическую разметку осуществили работники Института теоретической и компутационной лингвистики философского факультета Карлова университета в Праге (Гана Скоумалова, Милена Гнаткова, Томаш Елинек и Владимир Петкевич) в сотрудничестве с представителями Института формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге (Яном Гайичем и Ярославой Главачовой).

1. Сравнение частотности частей речи в избранных корпусах

В таблице 1 приводится частотность частей речи в КСЧ, SYN2005, ПРК, ЧСЧЯ и НС. Данные из КСЧ, SYN2005 и ПРК выделены жирным шрифтом, поскольку данные из SYN2005 и ПРК более надежны по сравнению с данными из ЧСЧЯ и НС. Таблица показывает, что между отдельными регистрами (художественный, публицистический, специальный, разговорный) существуют различия.

Заслуживает внимания сходство данных КЧС и SYN2005 (поэзия), причем не только в отдельных данных по частотности, но и в очередности частей речи (в таблице указывается цифрами в скобках). Корпусы отличаются частотностью употребления местоимений, числительных, предлогов, союзов и междометий, очередность же отлична лишь у числительных и междометий (КЧС в этом близок ПРК). Принимая во внимание тот факт, что КЧС исходит из текстов XIX века, а SYN2005 (поэзия) из текстов конца XX века, можно утверждать, что поэтический язык мало изменился в отношении частотности использования частей речи. Однако необходимо отметить, что мы сравниваем средние значения по КЧС и SYN2005 (поэзия), в то время как даже первый взгляд на данные по отдельным десятилетиям XIX века (как бы ни была проблематична такая периодизация) выявляет то,

Таблица 1

Частотность частей речи в избранных корпусах (в процентах)

	КЧС	**SYN 2005 (поэзия)**	**SYN 2005 (худ. проза)**	**SYN 2005 (спец. лит.)**	**SYN 2005 (публ.)**	**ПРК**	**ЧСЧЯ (поэзия)**	**ЧСЧЯ (худ. проза)**	**НС**
Имя сущ.	30,19 (1)	29,81 (1)	24,3 (1)	34,5 (1)	33,8 (1)	13,3 (4)	32,9 (1)	25,12 (1)	34,20 (1)
Имя прил.	9,91 (5)	10,06 (5)	8,9 (5–6)	15,4 (2)	12,2 (3)	5,2 (8)	9,96 (5)	8,78 (7)	19,08 (2)
Мест.	13,69 (3)	13,13 (3)	14,9 (3)	7,8 (5)	8,8 (5)	16,6 (2)	10,48 (4)	12,57 (3–4)	4,56 (7)
Имя числ.	0,64 (10)	1,07 (9)	1,6 (9)	3,3 (8)	3,3 (8)	0,5 (10)	1,32 (8)	1,71 (8)	1,15 (8)
Глагол	18,38 (2)	18,35 (2)	21,2 (2)	13,9 (3)	16,0 (2)	19,8 (1)	17,69 (2)	20,17 (2)	12,94 (3)
Наречие	7,75 (6)	7,96 (6)	8,4 (7)	5,5 (7)	6,2 (7)	8,5 (6)	8,98 (6)	10,97 (5)	8,33 (5)
Предлог	10,02 (4)	10,43 (4)	9,8 (4)	10,8 (4)	11,5 (4)	6,4 (7)	10,94 (3)	9,97 (6)	11,42 (4)
Союз	7,13 (7)	7,86 (7)	8,9 (5–6)	7,5 (6)	6,7 (6)	10,9 (5)	7,37 (7)	12,57 (3–4)	7,92 (6)
Частица	1,39 (8)	1,14 (8)	1,8 (8)	1,3 (9)	1,5 (9)	13,5 (3)	-	-	0,37 (9)
Междометие	0,89 (9)	0,2 (10)	0,11 (10)	0,02 (10)	0,02 (10)	1,5 (9)	0,36 (-)	0,26 (-)	0,03 (10)
количество единиц в корпусе	12856482	940573	40139930	32692293	39715768	548091	61087	487200	540000

что в некоторых случаях они достаточно заметно отличаются от средних значений по КЧС. Это соответствует высказыванию Мукаржовского о том, что «поэтический язык — это постоянная перемена» [Mukařovský 2007: 26].

Сравнение КЧС и SYN2005 (поэзия) с SYN2005 (художественная проза) показывает, что порядок частей речи в них в большей или меньшей степени соответствует. Заметнее оказывается разница в частотности союзов, что подтверждает наблюдение М. Р. Майеновой о большей степени асиндетичности стиха по сравнению с прозой [Mayenowa 1961]. В поэтических текстах не удивляет также и повышенная частотность междометий (уже отмечалось, что она приближается к крайним значениям разговорной речи). Кроме того, в поэзии наблюдается более высокий процент употребления именных частей речи (имена существительные, имена прилагательные), а в художественной прозе, соответственно, глагольных (глаголы, наречия); по всей вероятности, это связано с делением на эпический и лирический роды литературы, ср. пятый раздел настоящей статьи). В рамках четырех названных регистров SYN2005 (поэзия, художественная проза, специальный, публицистический) поэзия и проза близки друг другу, так же как регистры специальный и публицистический. Регистры специальной и художественной литературы (при включении ПРК это утверждение распространяется на специальный и разговорный регистры) также часто создают отдельные полюса (например, наибольшее количество имен существительных характерно для регистра специальной литературы, наименьшее — для художественной прозы, то же самое можно сказать об именах прилагательных и т. д.; напротив, больше всего наречий встречается в художественной прозе и меньше всего в специальных текстах). Так же рядом иногда оказываются и поэзия с публицистикой (разумеется, различия между поэзией и публицистикой заметнее, чем между поэзией и художественной прозой).

2. Частотность употребления частей речи в зависимости от длины стиха

Несмотря на то, что анжамбман (несовпадение синтаксического членения и стихораздела) относится к традиционным приемам и является одним из источников дифференциации стиля стиха, он даже в творчесте авторов, для которых считается характерным, затрагивает относительно небольшую долю стихов и отнюдь не ослабляет статистической значимости совпадения стихораздела и синтаксического членения [Červenka — Sgallová 1984: 13–14]. При переходе от коротких стихов к более длинным возрастает количество позиций, которые необходимо заполнить лексическими единицами. Следовательно, если действительно уравнение *граница стиха = граница синтаксического целого*, то для «заполнения» большего числа позиций в длинных стихах поэт может использовать de facto три возможности: 1) «заполнить» длинный стих бо́льшим количеством синтаксических единиц (предложениями), 2) использовать более длинные слова и/или 3) «заполнить» длинный стих второстепенными членами предложения. В настоящем разделе статьи мы сосредоточимся на ситуации в чешском стихе с этой точки зрения. При

этом мы будем заниматься не индивидуальными особенностями, а только общими тенденциями в исследуемом материале.

Ввиду того, что каждое предложение содержит только одну личную форму глагола (малочисленные исключения мы оставляем в стороне), наряду со средней длиной слова основными показателями для нас будут показатель $V_{FIN/СТИХ}$ (среднее количество личных форм в одном стихе), соответствующий среднему количеству предложений в одном стихе, и V_{FIN} (частотность употребления личных глагольных форм), т. е. косвенный показатель распространенности предложения.[3]

Таблица 2 / график 1

$V_{FIN/СТИХ}$, V_{FIN} и средняя длина слова в соответствии с длиной стиха

	6	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
$V_{FIN /СТИХ}$	0,59	0,66	0,74	0,8	0,88	0,96	1,01	1,1	1,09	1,17	1,25
V_{FIN}	17,1	16,6	16,28	15,78	15,71	15,34	15,29	15,08	14,67	14,62	15,07
Длина слова	1,56	1,68	1,69	1,7	1,71	1,72	1,73	1,74	1,83	1,82	1,86

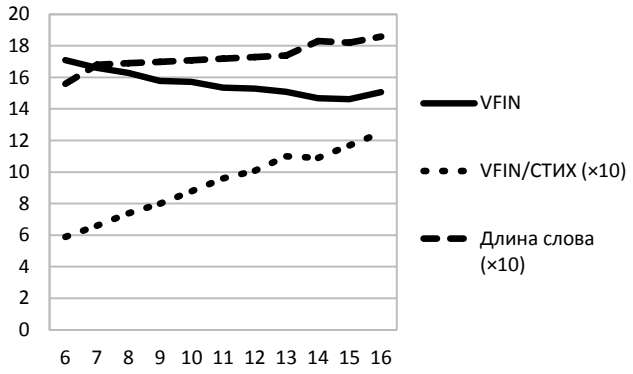

Таблица 2 и график 1 показывают, что в изучаемом материале встречаются все три упомянутые выше возможности. От 6-сложного стиха к 16-сложному средняя длина слова возрастает с 1,56 слога до 1,86 слога. Значения $V_{FIN/СТИХ}$ увеличиваются почти без колебаний с 0,59 до 1,25 личных форм в одном 16-сложном стихе. Таким

[3] V_{FIN} характеризует численность глаголов без учета форм, которые неспособны образовывать предложение (инфинитив, деепричастие, глагол *být* (быть) в условном наклонении, страдательное причастие). Проблематичными являются формы прошедшего времени, образуемые в чешском языке в 1-ом и 2-ом лице с помощью действительного причастия и формы глагола *být* (*četl jsem knihu* — я читал книгу), в 3-ем лице — при помощи одной только формы действительного причастия (*četl knihu* — он читал книгу). Эти две ситуации мы пока не можем отличить автоматически, поэтому в предложениях, содержащих 1-е или 2-е лицо прошедшего времени, при подсчете V_{FIN} засчитывались обе глагольные формы. Таким образом, показатель $V_{FIN/СТИХ}$ по сравнению с действительным соотношением предложение/стих немного преувеличен. Из-за этого могло появиться малозаметное искажение соотношений между лирическим (с предполагаемым бо́льшим количеством форм 1-го лица) и эпическим (предположительно бо́льшее количество форм 3-го лица) родами литературы. Тем не менее, численность включенных в V_{FIN} форм дает более точное представление о численности предложений, нежели обыкновенно используемое V.

образом, возрастание числа позиций ведет в поэтической речи, помимо прочего, к увеличению числа стихов, содержащих более чем одно предложение (или к понижению частотности предложений, разделенных на несколько стихов). Падение V_{FIN} (частотность личных форм глагола) одновременно указывает, что здесь также реализуется и третий способ, т. е. «заполнение» новых позиций второстепенными членами предложения. Следовательно, уместен вопрос, какие части речи служат для этого.

С частеречной точки зрения распространяющие члены предложения обычно являются именами прилагательными и наречиями. В соответствии с более ранними наблюдениями [Červenka — Sgallová 1984: 41], а также на основе нашего материала можно от 6-сложного к 16-сложному стиху проследить увеличение частотности имен прилагательных (8,63–11,26) и в заметно меньшей степени учащение употребления наречий (7,62–8,4). Однако более важным показателем, чем частотность употребления данных частей речи, является их количественное соотношение с личными формами глагола, т. е. прямые показатели распространенности предложения, а именно его именной (A/V_{FIN}) и глагольной (D/V_{FIN}) частей. Взаимное соотношение именной и глагольной частей может быть выражено коэффициентом AN/DV, т. е. соотношением количества имен прилагательных и существительных к количеству наречий и глаголов (в данном случае всех глагольных форм, не только личных)[4].

Таблица 3

Частотность частей речи

	6	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Имя сущ. (N)	29,7	29,59	29,27	29,1	29,16	29,29	29,89	29,24	30,56	29,41	29,32
Имя прил. (A)	8,63	9,16	9,05	9,43	9,59	9,64	10,25	10,06	11,29	11,26	11,2
Местоимение	14,7	14,69	14,57	14,68	14,7	14,73	14,19	14,2	13,45	13,71	13,93
Имя числ.	0,68	0,69	0,78	0,72	0,69	0,64	0,62	0,56	0,64	0,65	0,62
Глагол (V)	19,68	19,42	18,99	18,53	18,55	17,94	18,08	17,82	17,57	17,53	18,14
Наречие (D)	7,62	7,76	8,17	8,36	8,06	8,21	7,82	8,38	7,34	8,4	8,4
Предлог	10,76	10,41	10,41	10,16	10,37	10,21	10,53	10,16	10,33	10,05	10,31
Союз	6,37	6,45	6,84	7,06	7,13	7,53	6,95	7,8	7,28	7,3	6,53
Частица	1,33	1,36	1,42	1,44	1,34	1,35	1,26	1,29	1,15	1,33	1,21
Междометие	0,53	0,47	0,5	0,51	0,4	0,47	0,41	0,5	0,38	0,36	0,34

Таблица 4

Количественное соотношение употребления частей речи и личных форм глагола

	6	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
A/V_{FIN}	0,5	0,55	0,56	0,6	0,61	0,63	0,67	0,67	0,77	0,77	0,74
D/V_{FIN}	0,45	0,47	0,5	0,5	0,51	0,54	0,51	0,56	0,5	0,57	0,56
AN/DV	1,4	1,43	1,41	1,43	1,46	1,49	1,55	1,5	1,68	1,57	1,53

[4] В статье используются сокращения, принятые в морфологической разметке Чешского национального корпуса: A — имя прилагательное, D — наречие, N — имя существительное, V — глагол. — Прим. пер.

Итак, на основании данного критерия также можно наблюдать, что во всем исследуемом материале укрепляется тенденция распространения предложения именами прилагательными (непрерывный интенсивный рост A/V_{FIN} по сравнению с медленным и колеблющимся приростом D/V_{FIN}). Однако необходимо отметить, что приводившиеся до сих пор данные характеризуют чешский стих XIX века в целом, и что (как вскоре будет показано) именно разница в значениях этих параметров у сборников отдельных авторов (или «поэтических школ») и их отношение к значениям, полученным по целому КЧС, могут служить важными стилеметрическими показателями.

3. Частотность частей речи в зависимости от поэтической школы

Проблематику частотности употребления частей речи и ее зависимости от поэтической школы (или поколения) мы попробуем осветить, сравнивая значения, характеризующие творчество так называемых «майовцев» (50–60-е годы XIX века) и «люмировцев» (70–80-е годы XIX века) — см. таблицу 5[5].

Таблица 5

Частотность частей речи в зависимости от поэтической школы

	«Майовцы»	«Люмировцы»	КЧС
Имя существительное	29,71	30,12	30,19
Имя прилагательное	8,60	8,81	9,91
Местоимение	13,90	14,83	13,69
Имя числительное	0,67	0,70	0,64
Глагол	18,86	18,00	18,38
Наречие	8,08	8,20	7,75
Предлог	10,51	10,62	10,02
Союз	7,50	6,73	7,13
Частица	1,64	1,45	1,39
Междометие	0,52	0,53	0,89
V_{FIN}	16,59	15,43	15,77
$V_{FIN/СТИХ}$	0,80	0,76	0,92
A/V_{FIN}	0,52	0,58	0,66
D/V_{FIN}	0,49	0,53	0,49
AN/DV	1,42	1,49	1,53

Отличия между этими группировками невелики. У «люмировцев» немного выше значения у именной части предложения (имена существительные и прилагательные) и немного ниже у глагольной части (глаголы и наречия). Наряду с этим, отличаются значения у местоимений и союзов. Мы предполагаем, что объяснить данную разницу будет можно только по получении данных о метре стиха. Гораздо большие отличия в частотности частей речи были нами отмечены у отдельных авторов (см. следующий раздел).

4. Частотность частей речи в зависимости от авторского стиля

Проблематику частотности употребления частей речи и ее зависимости от авторского стиля мы попробовали осветить на материале произведений двух

[5] К «майовцам» мы отнесли Адольфа Гейдука, Витезслава Галека, Рудольфа Майера, Яна Неруду, Густава Пфлегера-Моравского и Вацлава Шольца; к «люмировцам» — Сватоплука Чеха, Йозефа-Вацлава Сладека, Ярослава Врхлицкого и Юлиуса Зейера.

Таблица 6

Разница в частотности употребления частей речи в творчестве Сватоплука Чеха и Виктора Дыка[6]

	Сватоплук Чех	**Виктор Дык**
Имя существительное	33,99	27,48
Имя прилагательное	12,74	8,99
Местоимение	12,60	14,63
Имя числительное	0,55	0,73
Глагол	14,91	23,06
Наречие	7,13	8,73
Предлог	10,99	8,23
Союз	5,31	6,26
Частица	1,29	1,42
Междометие	0,50	0,47
V_{FIN}	12,98	19,50
$V_{FIN/СТИХ}$	0,56	0,86
A/V_{FIN}	0,99	0,46
D/V_{FIN}	0,55	0,45
AN/DV	2,13	1,15

авторов — Сватоплука Чеха (1846–1908) и Виктора Дыка (1877–1931). На наш выбор повлиял тот факт, что стили этих авторов рассматриваются в литературоведении как противоположные друг другу. Процитируем высказывания о данных авторах, типичные в устах историков литературы. Я. Яначкова написала о Св. Чехе следующее: «Его пристрастие к описанию и апострофе, к искусно построенной фразе и витиеватому сложному предложению, благоприятствующему инверсивному порядку слов, напоминало современную ему политическую риторику и публицистику» [Lehár 1998: 315]. Й. Голы так характеризует особенности стиха В. Дыка: «Метрически правильный, немелодичный, без метафор, намеренно сухой, базируется на повторах, параллелизмах, подчеркнутых контрастах и непредсказуемых поворотах. При этом стих у Дыка часто совпадает с простым предложением» [Lehár 1998: 470]. Разница в частотности употребления частей речи в творчестве авторов достаточно значительна; это видно из таблицы 6.

Отличия затрагивают, прежде всего, соотношение именной (имена существительные и прилагательные) и глагольной (глаголы и наречия) частей предложения. Св. Чех отдает предпочтение первой группе, В. Дык, напротив, второй. Необходимо отметить также значения коэффициентов, характеризующих отношение имени прилагательного и наречия к личным формам глагола: по сравнению со стихом В. Дыка, стих Св. Чеха содержит гораздо больше распространяющих частей речи. Если мы вернемся к характеристике стиля Св. Чеха, то утверждение о его склонности к описанию можно подкрепить данными о возрастании количества имен прилагательных. Значения у В. Дыка отвечают констатации Й. Голого о том, что стиху В. Дыка всегда соответствует простое предложение (большое количество глаголов и личных глагольных форм, высокий коэффициент, указывающий на количество личных глагольных форм в одном стихе, меньшее число имен существительных и прилагательных). Согласно Я. Яначковой, стиль Св. Чеха напоминает характерную для его эпохи публицистику, сухой же, немелодичный стих В. Дыка приближается по этим своим характеристикам к прозе (как бы ни было поверхностно подобное сравнение). По этой причине уместно сравнивать данные о Св. Чехе с SYN2005 (публицистика), а данные о В. Дыке с SYN2005 (художественная проза) — см. таблицу 7. Для проверки приводим также средние данные из КЧС. Обращает на себя

[6] Приводятся средние значения, выведенные из значений по 4–12-сложным стихам.

внимание следующее: в тех случаях, когда значения у обоих авторов отличаются от КЧС, они зачастую приближаются или совпадают со значениями, соответствующими регистрам публицистики или художественной прозы. Вычисленные значения подтверждают различие авторских стилей с точки зрения частотности отдельных частей речи и в то же время указывают на то, что по некоторым характеристикам они отдаляются от поэтического регистра и приближаются к другим регистрам.

Таблица 7

Частотность употребления частей речи в творчестве Сватоплука Чеха и Виктора Дыка в сравнении с данными корпусов

	Св. Чех	SYN2005 (публицистика)	КЧС	SYN2005 (худ. проза)	В. Дык
Имя существительное	33,99	33,8	30,19	24,3	27,48
Имя прилагательное	12,74	12,2	9,91	8,9	8,99
Местоимение	12,60	8,8	13,69	14,9	14,63
Имя числительное	0,55	3,3	0,64	1,6	0,73
Глагол	14,91	16,0	18,38	21,2	23,06
Наречие	7,13	6,2	7,75	8,4	8,73
Предлог	10,99	11,5	10,02	9,8	8,23
Союз	5,31	6,7	7,13	8,9	6,26
Частица	1,29	1,5	1,39	1,8	1,42
Междометие	0,50	0,02	0,89	0,11	0,47

Таким образом, частотность употребления частей речи у Сватоплука Чеха в многочисленных случаях близка к показателям публицистики, в то время как у Виктора Дыка — к показателям художественной прозы.[7]

5. Частотность частей речи в зависимости от рода литературы

При описании проблематики частотности частей речи и её зависимости от рода литературы (лирического или эпического в данном случае) мы в настоящий момент ограничиваемся характеристикой творчества одного автора — Адольфа Гейдука (1835–1923).[8]

[7] Разумеется, это достаточно смелое утверждение, если учесть, что значения по художественной прозе и публицистике основываются на текстах конца XX века. Однако обнаруженное большое сходство между значениями по КЧС и SYN2005 (поэзия) дает возможность предполагать, что это сходство будет характерно и для прочих регистров.

[8] Данные об эпических произведениях А. Гейдука основываются на анализе произведений «Олдржих и Божена» („Oldřich a Božena“), «Завещание деда» („Dědův odkaz“), «Волынщик» („Dudák“), «На супрядках» („Na přástkách“), «Под Витковым камнем» („Pod Vítkovým kamenem“), Мехмед II („Mohamed II.“), «Дровосек» („Dřevorubec“), «За волю и веру» („Za volnost a víru“), «Бьела» („Běla“), «На волнах» („Na vlnách“); данные о лирике — на анализе произведений «Горечавка и пустырник» („Hořec a srdečník“), «В укромном месте» („V zátiší“) , «Стрелы и лучи» („Šípy a paprsky“), «Лесные цветы» („Lesní kvítí“), «На дорогах» („Na potulkách“), «Птичьи песни» („Ptačí motivy“).

Таблица 8

Частотность частей речи в лирических и эпических произведениях Адольфа Гейдука

	Лирический род	**Эпический род**	**Syn2005 (худ. проза)**	**КЧС**
Имя существительное	32,93	30,33	24,3	30,19
Имя прилагательное	9,96	8,50	8,9	9,91
Местоимение	12,98	13,53	14,9	13,69
Имя числительное	0,44	0,70	1,6	0,64
Глагол	17,18	19,42	21,2	18,38
Наречие	6,67	7,96	8,4	7,75
Предлог	10,65	11,18	9,8	10,02
Союз	7,42	6,50	8,9	7,13
Частица	1,29	1,52	1,8	1,39
Междометие	0,47	0,37	0,11	0,89

Как следует из табличных данных, различия невелики, однако можно заметить то, что предполагалось заранее: в лирике наблюдается б*о*льшая частотность именной части предложения, а в эпических произведениях — б*о*льшая частотность глагольной части[9]. Кроме того, таблица показывает, что значения эпического рода литературы по основным частям речи (имя существительное, имя прилагательное, местоимение, глагол, наречие) близки средним значениям в КЧС или значениям художественной прозы.

6. Заключение

1. Разница в частотности частей речи в КЧС (чешская поэзия XIX века) и в SYN2005 (подкорпус поэзии, т. е. оригинальных и переводных произведений конца XX века) малозаметна.

2. Сравнение корпусов или подкорпусов показывает, что между отдельными регистрами (художественный, публицистический, специальный, разговорный) существуют значительные различия.

3. Регистры специальный и разговорный (если не будет учитываться разговорный регистр, то на его месте окажется регистр художественной прозы) часто прямо противоположны. Регистр поэзии иногда приближается к регистру публицистическому, однако наибольшее сходство (даже несмотря на очевидные различия) он имеет с регистром художественной прозы.

4. Частотность частей речи зависит от длины стиха. Чем длиннее стих, тем больше предложений и второстепенных членов предложения (а значит, больше имен прилагательных и наречий) он включает. Именная часть предложения распространена более, чем глагольная, т. е. рост количества имен прилагательных заметнее, чем рост количества наречий.

[9] Ср.: «Эпический род литературы основывается на глаголе, лирический — на имени существительном или прилагательном». [Mistrík 1997: 519.]

5. Частотность частей речи зависит от поэтической школы. Эту зависимость будет необходимо изучить подробнее, поскольку она может объяснить сходства и различия там, где они неожиданны на основании традиционной литературоведческой характеристики.

6. Частотность частей речи зависит от конкретного автора. Отличия между отдельными авторами могут быть очень велики (вплоть до того, что некоторые значения приближаются к значениям, характерным для других регистров, например, для публицистики или художественной прозы).

7. Частотность частей речи зависит от рода литературы. В статье на примере анализа творчества одного автора было показано, что в лирическом роде более развитой оказывается именная часть (имена существительные и прилагательные), а в эпическом роде — глагольная часть (глаголы и наречия) предложения.

Литература

Гаспаров М.Л., Скулачева Т.В. Статьи о лингвистике стиха. М.: Языки славянской культуры, 2004. 284 с.

Bartoň T., Cvrček V., Čermák F., Jelínek T., Petkevič, V. Statistiky češtiny. Praha: NLN, 2009. 214 c.

Červenka M., Sgallová K. Český verš // Słowiańska metryka porównawcza II. Organizacja składniowa. Wrocław et al.: Zakład Narodowy im. Ossolińkich, 1984. C. 11–61.

Český národní korpus. Ústav Českého národního korpusu FF UK. Praha, 2010. [Электронный ресурс]. URL: http://ucnk.ff.cuni.cz

Ibrahim R., Plecháč P. Toward Automatic Analysis of Czech Verse // Barry P. Scherr, James Bailey, Evgeny V. Kazartsev (eds.). Formal Methods in Poetics. Lüdenscheid: RAM-Medienverlag, 2011. C. 295–305.

Jelínek J., Bečka J. V., Těšitelová M. Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha: SPN, 1961. 585 c.

Lehár J., Stich A., Janáčková J., Holý J. Česká literatura od počátků k dnešku. Praha: NLN, 1998. 1058 c.

Mayenowa M. R. Quelques différences entre un texte versifié et non-versifié (resumé) // Poetics — Poetyka — Поэтика. Warszawa: PWN, 1961. C. 369–371.

Mazáčová S. Sylabická délka věty v českém trocheji a jambu // Mayenowa M. R. (ed.). Semiotyka i struktura tekstu. Wrocław et al.: Zakład Narodowy im. Ossolińkich, 1973. C. 259–273.

Mistrík J. Štylistika slovenského jazyka. Bratislava: Slovenské pedagogické nakladateľstvo, 1997. 598 c.

Mukařovský J. O jazyce básnickém // Studie II. Brno: Host, 2007. C. 16–70.

Pszczołowska L. Długość wersu a budowa zdania // Mayenowa M. R. (ed.). Poetyka i matematyka. Warszawa: IBN, 1965. C. 79–96.

Šonková J. Morfologie mluvené češtiny: Frekvenční analýza. Praha: NLN, 2008. 356 c.

Těšitelová M. O básnickém jazyce z hlediska statistického // Slovo a slovesnost. Praha, 1968. C. 362–368.

Těšitelová M. Otázky lexikální statistiky. Praha: Academia, 1974. 289 c.

Těšitelová M. a kol. Kvantitativní charakteristiky současné češtiny. Praha: Academia, 1985. 249 c.

Těšitelová M. K současné české próze z hlediska frekvence slov // Naše řeč. Praha, 2000. C. 1–9.

Перевод с чешского Е. Маленинской

Robert Kolár, Petr Plecháč
Institute of Czech Literature, Czech Academy of Sciences
(Czech Republic, Prague)
kolar@ucl.cas.cz, plechac@ucl.cas.cz

THE FREQUENCY OF PARTS OF SPEECH IN CZECH POETRY

The study presents the first results from the statistical processing of the Czech Verse Corpus, which is a lemmatized, phonetically, morphologically, metrically, and strophically annotated corpus of Czech poetry from the 19th and the beginning of the 20th centuries. It contains 1700 books, with approximately 2.5 million verse lines and 15 million words. The paper compares the frequencies of parts of speech in the Czech Verse Corpus and other available corpora of Czech as well in their particular subcorpora (fiction, journalism, technical literature, spoken Czech). It also focuses on the relation between the frequency of parts of speech and the length of the line (the frequency of finite verbs, nominal vs. verbal phrases), literary movements (poetry of the 1850s and 1860s vs. poetry of 1870s vs. 1880s), author (Svatopluk Čech vs. Viktor Dyk) and literary genre (lyrics vs. the epics of Adolf Heyduk). As the study has been done on rather small samples, the results should be considered provisional. In the future we plan to broaden out material and to use the parameters analyzed here as stylometrical indicators that may be useful for both literary history and textual criticism.

Key words: versification, stylometry, corpus linguistics, parts of speech, poetics.

References

Bartoň T., Cvrček V., Čermák F., Jelínek T., Petkevič, V. *Statistiky češtiny* [Statistics on Czech]. Praha: NLN, 2009. 214 p. (In Czech)

Červenka M., Sgallová K. Český verš [Czech Verse]. *Słowiańska metryka porównawcza II. Organizacja składniowa* [Slavic Comparative Metrics II. Syntax]. Wrocław et al.: Zakład Narodowy im. Ossolińkich, 1984. Pp. 11–61. (In Czech)

Český národní korpus [Czech National Corpus]. *Ústav Českého národního korpusu FF UK*. Praha, 2010. Available at: http://ucnk.ff.cuni.cz, accessed 11.05.2017. (In Czech)

Gasparov M. L., Skulacheva T. V. *Stat'i o lingvistike stikha* [Studies on Linguistics of Verse]. M.: Yaziki slavianskoi kultury, 2004. 284 p. (In Russ.)

Ibrahim R., Plecháč P. Toward Automatic Analysis of Czech Verse. Barry P. Scherr, James Bailey, Evgeny V. Kazartsev (eds.). *Formal Methods in Poetics*. Lüdenscheid: RAM-Mediencerlag, 2011. Pp. 295–305. (In Engl.)

Jelínek J., Bečka J. V., Těšitelová M. *Frekvence slov, slovních druhů a tvarů v českém jazyce* [Frequencies of Words, Parts of Speech, and Word Types in Czech]. Praha: SPN, 1961. 585 p. (In Czech)

Lehár J., Stich A., Janáčková J., Holý J. *Česká literatura od počátků k dnešku* [Czech Literature from the Beginnings to Today]. Praha: NLN, 1998. 1058 p. (In Czech)

Mayenowa M. R. Quelques différences entre un texte versifié et non-versifié (resumé) [On Some Differences Between Versified and Non-Versified Texts (summary)]. *Poetics — Poetyka — Поэтика*. Warszawa: PWN, 1961. Pp. 369–371. (In French)

Mazáčová S. Sylabická délka věty v českém trocheji a jambu [Sentence Length in the Czech Trochee and Iamb Measured by the Number of Syllables]. Mayenowa M. R. (ed.). *Semiotyka i struktura tekstu* [Semiotics and Structure of a Text]. Wrocław et al.: Zakład Narodowy im. Ossolińkich, 1973. Pp. 259–273. (In Czech)

Mistrík J. *Štylistika* [Stylistics]. Bratislava: Slovenské pedagogické nakladateľstvo, 1997. 598 p. (In Slovak)

Mukařovský J. O jazyce básnickém [On Poetic Language]. *Studie II* [Papers II]. Brno: Host, 2007. Pp. 16–70. (In Czech)

Pszczołowska L. Długość wersu a budowa zdania [Length of a Verse Line and Sentence Construction]. Mayenowa M. R. (ed.). *Poetyka i matematyka* [Poetics and Mathematics]. Warszawa: IBN, 1965. Pp. 79–96. (In Polish)

Šonková J. *Morfologie mluvené češtiny: Frekvenční analýza* [Morphology of Spoken Czech: Frequency Analysis]. Praha: NLN, 2008. 356 p. (In Czech)

Těšitelová M. O básnickém jazyce z hlediska statistického [On Czech Language From the Point of View of Statistics]. *Slovo a slovesnost* [The Word and Verbal Culture]. Praha, 1968. Pp. 362–368. (In Czech)

Těšitelová M. *Otázky lexikální statistiky* [Problems of Lexical Statistics]. Praha: Academia, 1974. 289 p. (In Czech)

Těšitelová M. a kol. *Kvantitativní charakteristiky současné češtiny* [Quantitative Characteristics of Contemporary Czech]. Praha: Academia, 1985. 249 p. (In Czech)

Těšitelová M. K současné české próze z hlediska frekvence slov [On Contemporary Fiction with Regard to Word Frequencies]. *Naše řeč* [Our Speech]. Praha, 2000. Pp. 1–9. (In Czech)