# Инструкция по выявлению и анализу предложных конструкций

### Вступление

Полученные данные для каждого предлога следует сохранять в папке с именем, совпадающим с этим предлогом и вашей фамилией через подчерк. В этой папке должны содержаться следующие файлы:

- 1) ВАШ ПРЕДЛОГ\_Report
- 2) ВАШ ПРЕДЛОГ\_Статистика\_Корпус (NB: для разных корпусов может быть несколько файлов или несколько листов в файле Excel)
- 3) ВАШ ПРЕДЛОГ\_Конструкции\_Корпус (NB: для разных корпусов может быть несколько файлов или несколько листов в файле Excel)
- 4) ВАШ ПРЕДЛОГ\_Сводная
- 5) ВАШ ПРЕДЛОГ\_Значения по 3 словарям и Синонимы

## Часть 1.

1. Корпус: Araneum Russicum Minus 120 М

 Сформировать конкорданс сочетаний «ХОЗЯИН+ваш ПРЕДЛОГ+СЛУГА»: Search: Query type: CQL.

Например, для предлога «для» запрос CQL я бы составил так:

```
[atag != "Zz"]
[atag = "Vb"]
[atag != "Zz"& atag != "Vb"] {0,2}
[lemma="для"]
[atag = "Aj"|atag = "Pn"]{0,2}
[atag="Nn"]
[atag = "Nn"]
[lemma="для"]
[atag = "Aj"|atag = "Pn"]{0,2}
[atag="Nn"]
[atag!="Zz"]
```

## (для любых корпусов Aranea

(вместо русского корпуса и предлога могут быть любые другие))

#### или

```
["[[:punct:]]*"]
[tag="V.*"]
[word != "[[:punct:]]*"& tag != "V.*"] {0,2}
[lemma="для"]
[atag = "Aj"|atag = "Pn"]{0,2}
[atag="Nn"]
[lemma="для"]
[atag = "Nn"]
[atag = "Aj"|atag = "Pn"]{0,2}
[atag="Nn"]
["[[:punct:]]*"]
```

(для любых русских корпусов как в Aranea, так и в системе Sketch Engine)

Запрос запрашивает формирование конкорданса со следующими конструкциями: Существительное или глагол, потом 1 или 2 слова, не являющиеся знаком препинания, существительным или глаголом, или ни одного, потом «для», потом 1 прилагательное или местоимение или ни одного слова, потом существительное

Возможно другое количество слов {0,К} между хозяином и предлогом, предлогом и слугой... Возможен более сложный запрос, когда слово-хозяин находится справа.

NB: кроме хозяина-существительного или глагола предлог может управляться и прилагательным (причастием), кроме слуги-существительного предлог может управлять местоимением или аббревиатурой, здесь это не описано. Чтобы это учесть, надо чуть переделать запрос. Но при этом возникает вероятность «шума». Возможно, эти случаи лучше изучить на основе отдельных запросов.

NB: Важно: Исключить из результатов поиска для первообразных предлогов случаи, где первообразный предлог является частью производного составного, напр.: в виду, в связи, с целью, вблизи от, для ради... Для этого подобные конструкции надо описать в запросе с отрицанием, напр., вышеуказанный запрос надо модифицировать так:

......[lemma="для"][word!="<mark>ради</mark>" & (atag = "Aj"|atag = Pn")]{0,2}[atag="Nn"]

чтобы исключить контексты с предлогом «для ради» («Во времена СССР лгали всему миру для ради своей выгоды»)

Если вторая часть составного предлога является существительным, то надо исключить ее появление в позиции «слуги»: *word!="виду", а именно:* 

[lemma="в"|lemma="во"][atag = "Aj"|atag = "Pn"]{0,2}[atag="Nn" & word!="виду" & word! ="целях" & word!="имя"]

Производные предлоги-синонимы - см. файл «Производные предлоги Захаров».

3. Обработать полученные данные:

3.1а. Сохранить запрос в файле ВАШ ПРЕДЛОГ\_Report (см. Приложение 1).

3.16. Сохранить (записать) статистический результат поиска: Query ...... XXXXX (ZZ.ZZ per million) в файле ВАШ ПРЕДЛОГ\_Report (Приложение 1).

NB: В начале этого файла записать фамилию-имя исполнителя и дату формирования файла. Там же писать любые вопросы-замечания по выполнению этой работы.

3.2. Sample: 2000 (случайная выборка, см. Приложение 2).

3.3. Frequency: first level – lemma – Node (см. Приложение 3).

3.4. Save: 2000 lines (см. Приложения 4, 5). Сохраненный список конструкций переименовать в «ВАШ ПРЕДЛОГ\_Конструкции\_Корпус».

### ВОПРОС: в какой программе, в каком виде и формате лучше хранить конструкции?

3.5. Из полученного списка (lemma, frequency), например:

lemma Frequency

маска для волос 6 товар для дом 5 оборудование для производство 4 место для отдых 4 средство для уход 3 средство для похудение 3 предназначить для работа 3

#### выбрать первые первые 100 строк и проанализировать их.

При попадании в результат плохих (неправильных) конструкций или одинаковых конструкций разной длины (возникающих, в частности. из-за присутствия в запросе элементов типа []{0,2}), напр., «учебник для школа» и «учебник для средний школа») и т. д. **плохие (лишние) конструкции** удалить и дополнить список до 100 случайными конструкциями из сохраненного списка 2000 конструкций.

Как выбрать случайно нужное количество конструкций из оставшихся 1900, чтобы дополнить список до 100? Для этого в Excel можно добавить столбец "случайное число" при помощи фунции =СЛЧИС(), упорядочить эту часть списка по возрастанию или убыванию чисел в этом столбце и выбрать те, которые дополнят наш список до сотни.

3.6. Затем перекинуть эти 100 конструкций в Excel (при этом разделительную линию из дефисов убрать). К этой таблице добавить еще 5 столбцов: «Значение», «Процент», «Суммарная частота», «Процент по случайной выборке» и «ipm».

4. Просмотреть все сочетания, определить **значения предлога** в каждом из них (См. описание своего предлога с примерами – файл *Онтологическое описание простых непроизводных предлогов*) и вписать его в столбец «Значение».

NB: В ряде случаев для определения значения рекомендуется обращаться к расширенному контексту, а том числе через Frequency list. Там, где нужен более широкий контекст, копировать его в таблицу в последний столбец «Комментарии». NB: Там, где ни одно из значений «онтологического описания не подходит, добавлять свое с комментарием. Одним из таких неучтенных вариантов может особое фразеологизированное значение.

5. По завершении просмотра и определения значений предлога в каждой конструкции:

5.1. отсортировать таблицу по значениям предлогов (3 столбец), и по частоте (2 столбец),

5.2. посчитать процент конструкций с данным значением и поместить его в ячейку в столбец «Процент» (если обрабатываем 100 строк-конструкций, то это просто количество конструкций с данным значением),

5.3. суммировать все количества для одинаковых значений предлога из столбца «Frequency» (которые после сортировки соберутся рядом) и поместить сумму в столбец «Суммарная частота» (в первую строку с данным значением),

5.4. посчитать процент для каждого значения (1 знак после запятой) и поместить его в столбец «Процент по случайной выборке» в соответствующую строку,

5.5. Посчитать ipm для каждого значения (2 знака после запятой) (процент от значения ipm в п. 3.1б) и поместить его в столбец «ipm» в соответствующую строку.

6. Сохранить таблицу с именем «ВАШ ПРЕДЛОГ\_ Статистика\_Корпус» (см. Приложение 6).

7. Отдельно делаем сводную таблицу (на отдельном листе) с частотами по значениям (без конструкций).

#### Часть 2.

Собираем такую же статистику для своего первообразного предлога по другим корпусам, в данный момент, НКРЯ «Основной корпус» (100 случ. контекстов) и НКРЯ «Газетный корпус» (100 случ. контекстов).

NB: НКРЯ: Конкорданс КWIC, документов на странице 100, упорядочить «случайно».

NB: Для НКРЯ подробно технологический процесс не описан. Желающих поделиться опытом, просим описать его и послать по почте (vz1311@yandex.ru) или поместить в файле ВАШ ПРЕДЛОГ\_Report.

NB: В силу трудоемкости работы с НКРЯ сохраняемую случайную выборку предлагаем формировать в объеме 500 конструкций.

### Часть 3.

Собираем статистику по **производным предлогам-синонимам** по 3 выше названным корпусам. Как правило, это частота в корпусе и ipm.

Но если производный предлог имеет несколько значений, то кроме частоты для каждого значения посчитать процент (по технологии как для первообразных предлогов) и соответствующую ему долю для ipm.

Если синоним является первообразным предлогом, то для него дать только суммарную информацию (частота и ipm) - процент для соответствующего значения мы возьмем из его основного файла. Производные предлоги-синонимы - см. файл «Производные предлоги Захаров» (пока чер</mark>новой вариант).

Также можно искать: в голове, в толковых словарях, в Вики-словаре.

Часть 4.

Корпус SynTagRus и другие корпусы

ПОКА ПРОПУСКАЕМ!

Часть 5.

Свести все данные в одну таблицу – см. как образец файл ПОД\_Сводная\_Плетнева.docx (часть сводной таблицы см. в Приложении 7).

#### Файл ПРЕДЛОГ\_Report (можно его расширить, писать комментарии, вопросы и т.п.):

**Предлог ДЛЯ** <u>Исполнитель</u>: Захаров В.П. <u>Дата</u>: 23 августа 2018 г. <u>Корпус</u>: Araneum Russicum II Minus (Russian, 17.03) 120 M <u>Запрос</u>: [tag="N.\*"]tag="V.\*"] [word != "[[:punct:]]\*"& tag != "N.\*"& tag != "V.\*"] {0,2} [lemma="для"][ tag = "A.\*"|tag = "P.\*"]{0,1} [tag = "N.\*"] <u>Результат поиска</u>: Query **N.\***, **V.\***, **[[:punct:]]\***, **для**, **A.\***, **P.\* 370**, **435** (3,087.00 per million) <u>Случайная выборка</u>: Query N.\*, V.\*, [[:punct:]]\*, для, A.\*, P.\* **370**, 435 > Random sample 1,000 (8.33 per million)

#### Приложение 2

Home	Ourse Mar Mit	fferentillt and man At Dr. off and a Deadland		
Search	Query Nn, Vo,	[[:punct:]]", для, ради, АJ, Рп 300,117 > Капоог	n sample 2,000 (16.6/ per million)	
Word list	Page 1	of 100 Go Next   Last		
Corpus info	rinco.kz	или чугуна; ¶ стиральный порошок не лучший	материал для чистки	кабинки; ¶не реж
Myjobs	bohemia.kz	печь. ¶Формы Silikomart – это отличный	подарок для любителей	побаловать себя и
My jobs	wac.uz	от бизнес-плана следующим: ¶обычно ТЭО	пишется для проектов	внедрения новых
User guide 년	irkipedia	. Были проведены беспрецедентные инженерные	изыскания для отвода	воды. Была создал
	broker.ru	на поверхности. Слушатели получат простые	рекомендации для ответов	на сложные вопро
Save	checker.ru	исходящих,и вопросы пропали сами собой. f	Юрий для МТС	: 20.08.11 22:36. м
View options	kpkmedic.r	и объемы тела – комфортно, без операции, без	диет для быстрого похудения	и без отрыва от пլ
KWIC	club.shopo	. И сделаем это на примере приобретения (	струн Elixir для электрической бас-гитарь	и. Как лучше всего
Sentence	waithaispa	и сертификатам. Каждое мое посещение-это	польза для здоровье	и душевный отды
Sort	ibeauty-he	многих городах страны сильно загрязнена, в ней	находятся опасные для здоровья	человека токсины
Left	sksauto.ru	и на руках будут необходимые средства. В целях	безопасности и для удобства	, можно оформиті
Right	brimz.ru	личность. $\P \equiv \mathbf{C}$ того момента как я начал делать	упражнения для глаз	прошло 2 недели.
Node	ps-imho.co	другую должность, можно выделить специальное	время для общения	друг с другом, ког
Poferoneos	ngcms.ru	минусы. ¶В любом случае статья будет выложена в	форум для оценки	и внесения измен
Chuffle	pressria.r	по приглашению ведущих ученых из других	страх для чтения	конкретных лекці
Shune	angara-gbi	выдерживают высокие несущие нагрузки и	применяются для строительных работ	любой сложности
Sample	visardo.ru	Федерацию. ¶ Виза в Россию. Оформление	приглашений для иностранцев	в Россию. ¶Как п <u>і</u>
Last (2000)	visardo.ru	превышать 28 дней. Полное оформление данного	приглашения для иностранца	займет около одн
Filter	rusfolder	во время регистрации. На него будет выслана	ссылка для изменения	пароля. При пере
Sub-hits	ristikivi	дни и ночи в ожидании близкого конца. ¶Все	хлопоты для получения	пропуска через гр
1st hit in doc	Page 1	of 100 Go Next   Last		
Frequency				

Приложение 3

Home	English and Bat		
Search	Frequency list		
Word list	Frequency limit: 0 Set limit		
Corpus info			
Myjobs	Page 4 C. Next		
User guide	Page 1 Go <u>Next&gt;</u>		
eser galace	word	<b>Frequency</b>	Items: 1,981    Total frequency: 2,000
Care	<u>Р</u>   <u>N</u> препаратов для лечения	3	
Save	<u>Р</u>   <u>N</u> устройств для поддержания	2	
<b>O</b> concordance	<u>Р</u>   <u>N</u> условия для обучения	2	
Frequency	<u>Р</u>   <u>N</u> товаров для животных	2	
Node tags	<u>Р</u>   <u>N</u> способ для заработка	2	
Node forms	<u>Р</u>   <u>N</u> ресурсы для крафта	2	
Dee IDe	<u>Р</u>   <u>N</u> применяется для профилактики	2	
Docids	<u>Р</u>   <u>N</u> предназначены для питания	2	
0	<u>Р</u>   <u>N</u> предназначено для создания	2	
	<u>Р</u>   <u>N</u> подарок для любителей	2	
	<u>Р</u>   <u>N</u> пароль для входа	2	
	<u>Р</u>   <u>N</u> молочко для тела	2	
	<u>Р</u>   <u>N</u> место для отдыха	2	
	<u>Р</u>   <u>N</u> мебель для спальни	2	
	<u>Р</u>   <u>N</u> крем для депиляции	2	
	<u>Р</u>   <u>N</u> инфраструктуру для проведения	2	
	<u>Р</u>   <u>N</u> включен JavaScript для просмотра	2	10 m
	<u>Р</u>   <u>N</u> вещи для малыша	2	No. of the second se
	<u>Р</u>   <u>N</u> ящиками для кухонной утвари	1	
	<u>Р</u>   <u>N</u> яркости для определения	1	
	<u>Р</u>   <u>N</u> языку как для общеобразовательных школ	1	
	<u>Р</u>   <u>N</u> языка для школьников	1	
	<u>Р</u>   <u>N</u> языка для взрослых	1	line and the second sec
	<u>Р</u>   <u>N</u> язык для европейца	1	
	<u>Р</u>   <u>N</u> являются фундаментальными для понимания	1	

#### Приложение 4



1

Приложение 5

#### Файл ПРЕДЛОГ\_Конструкции\_Корпус

word Frequency
препаратов для лечения 3
устройств для поддержания 2
условия для обучения 2
товаров для животных 2
способ для заработка 2
ресурсы для крафта 2
применяется для профилактики 2
предназначены для питания 2
предназначено для создания 2
подарок для любителей 2
пароль для входа 2
молочко для тела 2
место для отдыха 2
мебель для спальни 2
крем для депиляции 2
инфраструктуру для проведения 2
включен JavaScript для просмотра 2
вещи для малыша 2
ящиками для кухонной утвари 1
яркости для определения 1
языку как для общеобразовательных школ

# Приложение 6

# Файл ПРЕДЛОГ\_Статистика\_Корпус

# (NB: в этом примере обработано 200 конструкций)

	А	В	С	D	E	F	
1	lemma	Frequency	Значение	Суммарная ча	Процент	ipm	
2	подходить для весь тип	3	Авторизатор	11	5,8	179,05	
3	год для мужчина	2	Авторизатор				
4	быть для человек	2	Авторизатор				
5	язык для студент	1	Авторизатор				
6	язык для дитя	1	Авторизатор				
7	являться для местный житель	1	Авторизатор				
8	эпоха для московский университе	1	Авторизатор				
9	являться идеальный и для хозяин	1	Носитель признак	1	0,4	1,23	
10	маска для волос	6	Финдест.	177	93,6	2339,43	
11	товар для дом	5	Финдест.				
12	оборудование для производство	4	Финдест.				
13	место для отдых	4	Финдест.				
14	средство для уход	3	Финдест.				
15	средство для похудение	3	Финдест.				
16	предназначить для работа	3	Финдест.				
17	подходить для создание	3	Финдест.				
18	площадка для игра	3	Финдест.				
19	оборудование для проведение	3	Финдест.				
20	мебель для кухня	3	Финдест.				
21	участок для строительство	2	Финдест.				
22							

# Файл ПРЕДЛОГ\_Сводная

# <u>(NB: Это пример задания по старой инструкции, отличающейся от данной - немного</u> другая методика и обрабатывалось 200 или больше конструкций, в самой таблице возможны ошибки — дается только как образец)

предлог¶	значение¶	НКРЯ со снятой омон. 6 млн. слов ← (200 случ. контекстов)¶		НКРЯ газетный, 228,5-млнслов (208 случконтекстов)¶		Агапеит Russicum веб- корпус-120 млн.¶ (200 случ. контекстов/ ← 200 частотных контекстов/)¶		Комментарии¶	
		употр.	ipm¶	употр.	ipm	употр.	ipm¶	9	
nod¶	директив¶	٢	1	٩	٩	5·(2,51%)/↔ 180·(2,11%)¶	19,9 26,9	٩	
близ¶	директив	٢	4	\$	٢	1013	8,4	[lemma="близ ".&. tag="Sp.*".]. (для. омонимичных предлогов)	
возле	директив	٩	•	e	9	5 (2,5%)¶	1,19	ſ	
возле¶	ДРУГОЕ ЗНА ЧЕНИЕ не из списка значений для	٩	4	4	ę	9 (4,75%)	2,19	4	
08070	лиректив	C	C	¢	C	73 (36.5%)	92.1	¢	
BOKDV2	лиректив	C	C	C	C	9800	81.7	¢	
рядом с	лиректив	¢	<	C.	¢	27 (13.5%)	7.9	ſ	
справа-от	директив	ſ	•	C	e.	381	3.2	¢	
слева от¶	лиректив	¢	¢	¢	¢	350	2.9	¢	
со стороны	лиректив	(	•	•	¢	10712	89.3	< Comparison of the second sec	
сбоку.om	лиректив	¢	•	¢	¢	69	0.57	¢	
поблизости от	директив	(	<	•	¢	17.(14%)	0.14	ſ	
no•nvmu¶	директив	ſ	•	C	C I	1557	13	¢	
nod¶	дестинатив	٩	٩	٩	٩	36 (18,09%) 754 (8,84%)	143,2¶ 122,9¶	1	
с-целью ¶	дестинатив	•	9	٩	9	8227	68,6	٩	
nod¶	способ <sup>.</sup> действия¶	٩	٩	٩	٩	7·(3,52%)¶ 25·(0,29%)¶	43,2 3,7	٢	
с-помощью¶	способ действия	٩	٩	٩	٩	163 (81,5%)	186,3¶	٩	
при-помощи¶	способ действия	1	٩	٩	٩	186 (93%)	76,7¶	٩	