Инструкция по выполнению семантического анализа значения предлогов на базе текстов Национального корпуса русского языка

Программные средства. Средства поиска и представления данных НКРЯ. Текстовый процессор MS Word. Электронная таблица MS Excel.

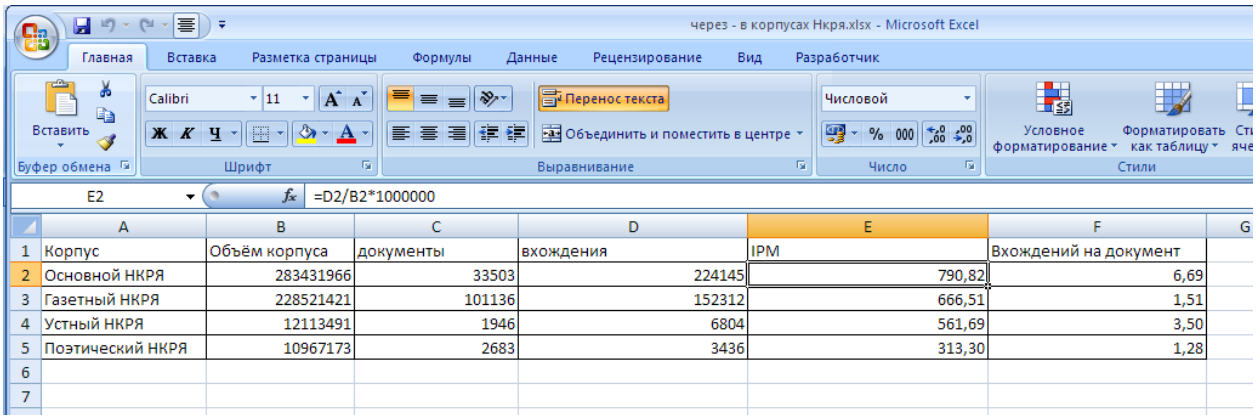
Ход работы.

А) Исследование частотных характеристик предлогов

1. Произвести по поиск в основном корпусе НКРЯ по исследуемому предлогу.
2. Зафиксировать число документов, число вхождений и объём корпуса.
3. Используя функцию «поискать в других корпусах» выполнить поиск в газетном, устном и поэтическом корпусах
4. Результаты оформить в таблице следующего вида

Корпус НКРЯ	объём корпуса	документы	вхождения	IPM	Вхождений на документ

5. Для подсчета IPM и параметра «вхождений число вхождений на документ таблицу перенести в MS Excel
6. В окно «вставить функцию» вставить формулу, как показано на рис. 1



	A	B	C	D	E	F	G
1	Корпус	Объём корпуса	документы	вхождения	IPM	Вхождений на документ	
2	Основной НКРЯ	283431966	33503	224145	790,82	6,69	
3	Газетный НКРЯ	228521421	101136	152312	666,51	1,51	
4	Устный НКРЯ	12113491	1946	6804	561,69	3,50	
5	Поэтический НКРЯ	10967173	2683	3436	313,30	1,28	
6							
7							

Рис. 1. Таблица «Частотные характеристики предлога «через» в текстах корпусов НКРЯ»

7. IPM рассчитывается по формуле: число вхождений (в таблице ячейка D2) / объём корпуса (в таблице ячейка B2) * 1000000. Значение число вхождений на документ – подсчитывается делением числа вхождений на число документов. Для ячеек выбрать числовой формат с двумя знаками после запятой.

Б) Количественный анализ значений предлогов

Произвести поиск в корпусе НКРЯ по исследуемому предлогу

Задать настройки представления результатов как показано на рис 2

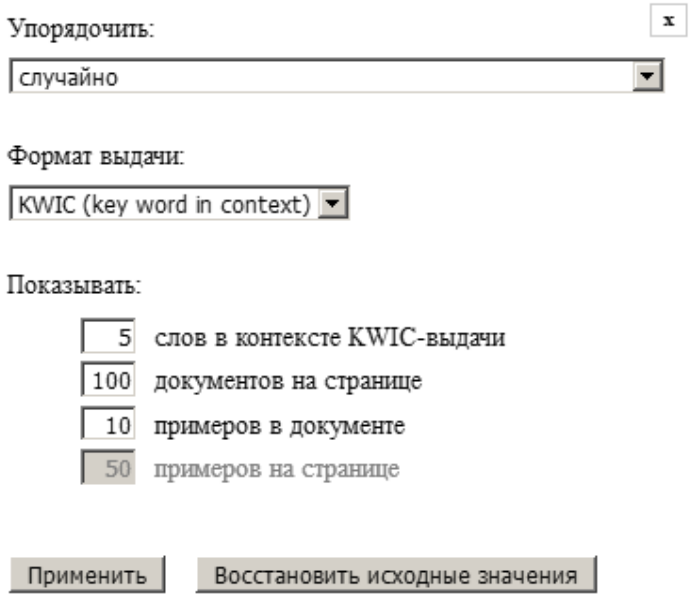


Рис 2. Настройки представления результатов поиска

Результаты на странице выделить с помощью мыши и перенести в файл MS Word, где они будут иметь вид сложной таблицы с большим числом вставных таблиц и ссылок, как показано на рис. 3.

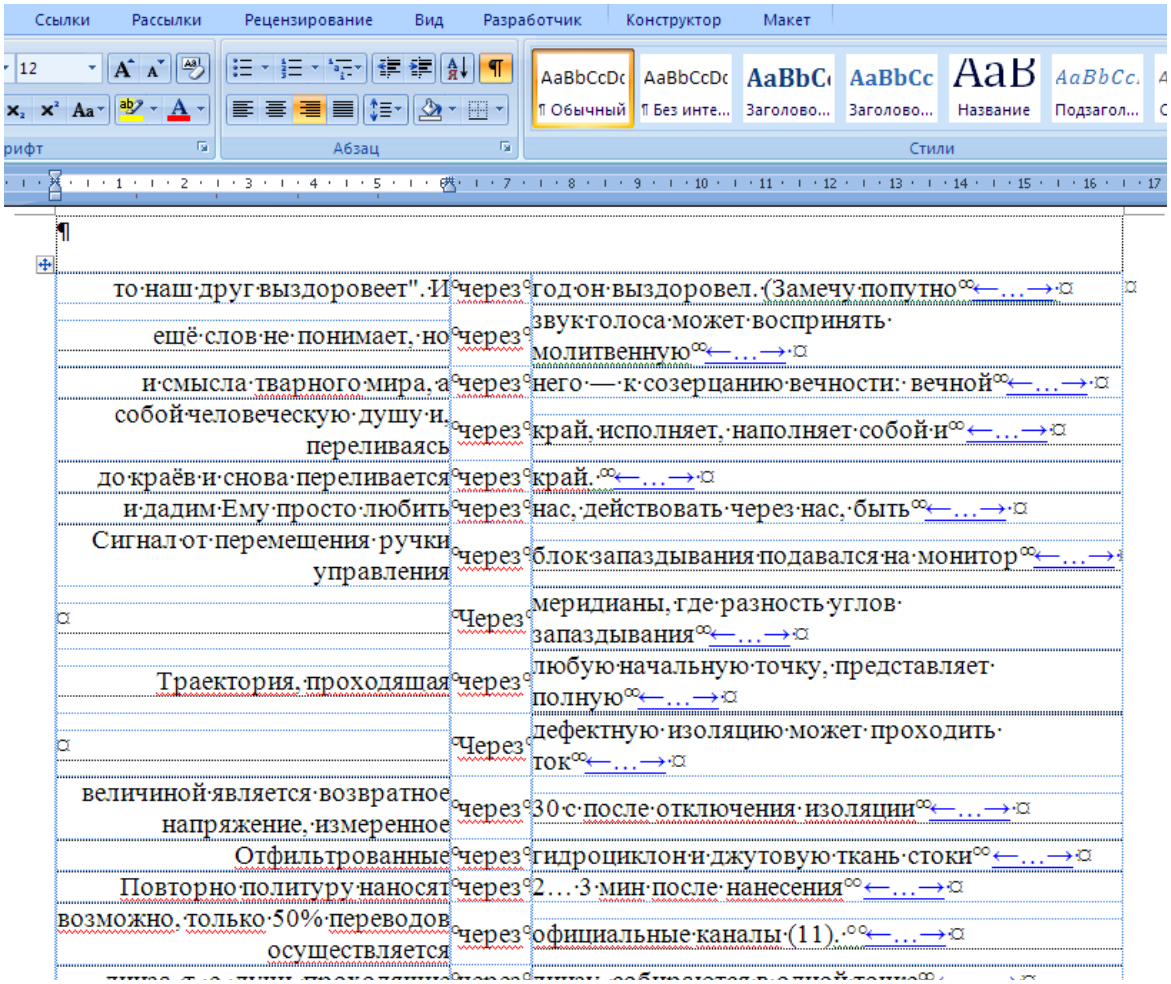


Рис. 3 Результаты поиска в НКРЯ после перенесения их в MS WORD

В таком представлении данные не поддаются автоматической нумерации, практически невозможно добавить необходимые для ввода дополнительных данных строки и столбцы. В связи с этим необходимо преобразовать таблицу в более удобный формат. Для этого следует выделить таблицу. Выделять таблицу следует с помощью мыши, так как функция «выделить таблицу» Word не работает. Затем следует использовать функцию «преобразовать в текст», задав при этом следующие параметры (рис. 4)

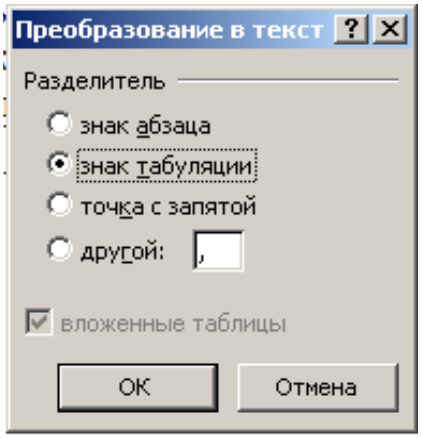


Рис. 4 Параметры преобразования таблицы в текст

Полученный результат будет иметь следующий вид (рис. 5)

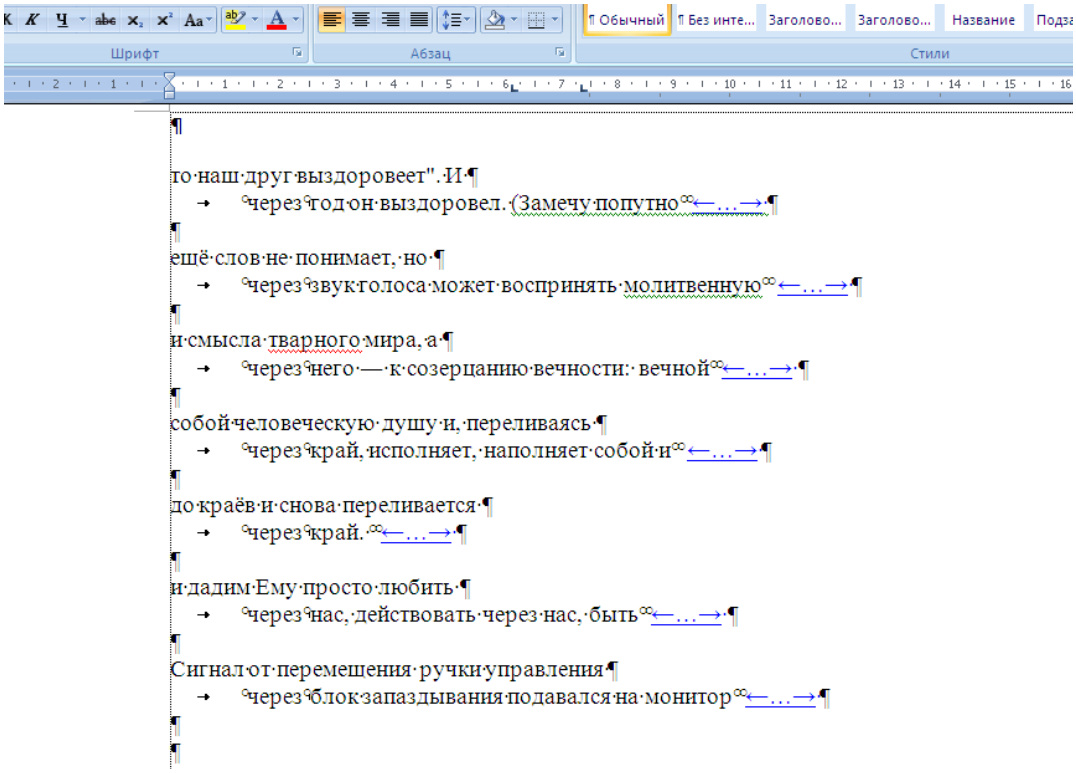


Рис. 5 Вид данных после преобразования таблицы в текст

Далее следует выполнить несколько контекстных замен.

Знак абзаца – знак табуляции на пробел

Знак табуляции на пробел

Неразрывный пробел на удаление

Знак гиперссылки на удаление (для этого выделить гиперссылку и открыть функцию «заменить»)

Два пробела на один

Два знак абзаца на один

Результат – текст в формате Word разбитый по абзацам, каждый из которых представляет собой контекст использования предлога. Рекомендуется с помощью контекстной замены задать цвет шрифта исследуемого предлога (рис. 6), что несколько упростит дальнейшую работу.

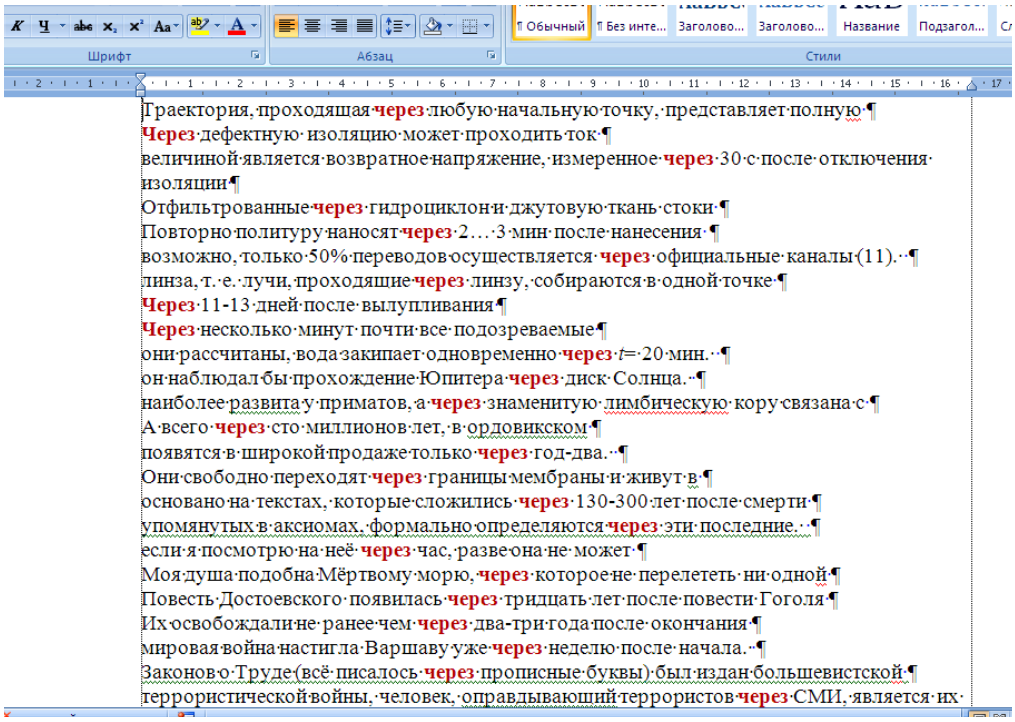
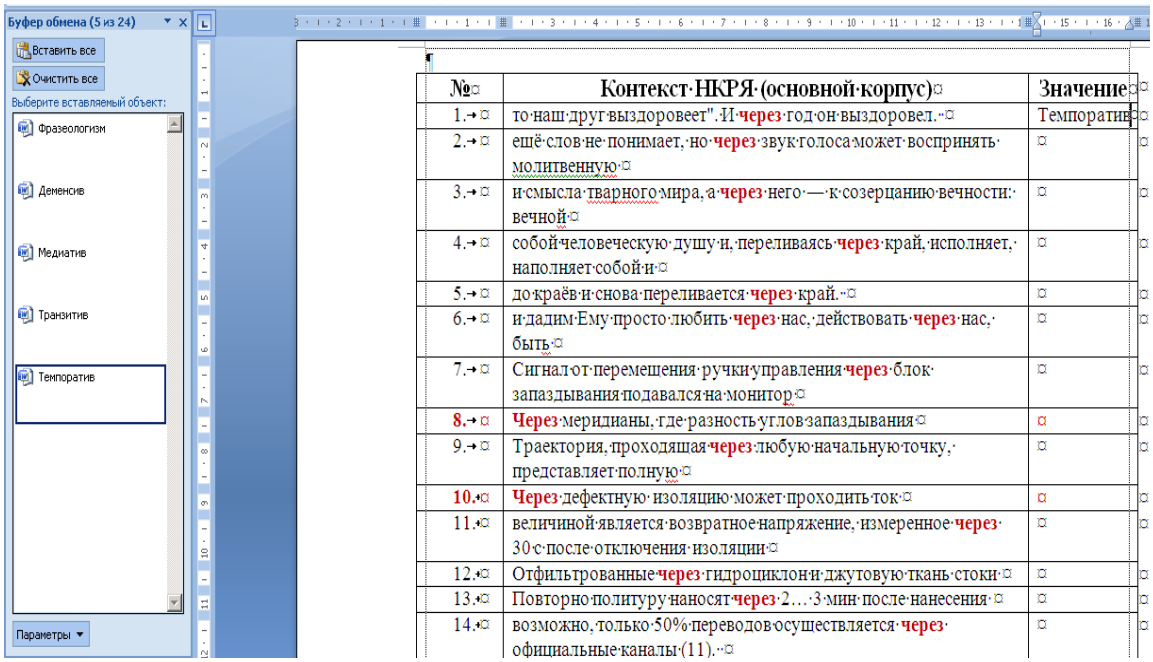


Рис. 6. Промежуточный результат преобразования исходных данных.

Примечание. Если необходимо обработать несколько страниц результатов поиска в НКРЯ, то следующую страницу рекомендуется загрузить в тот же файл, после завершения данного этапа. С ней производятся описанные выше операции, затем тексты соединяются в один.

Текст выделить и через меню «вставить таблицу» преобразовать его в таблицу с начала с одним столбцом, а за тем добавить ещё два – слева и справа (рис. 7). Столбцы озаглавить

«№», «Контекст НКРЯ (Основной корпус)», «Значение». Открыв буфер обмена, в него следует загрузить наименования значений (рис 7).



№	Контекст НКРЯ (основной корпус)	Значение
1.	то наш друг выздоровеет". И через год он выздоровел.	Темпоратив
2.	ещё слов не понимает, но через звук голоса может воспринять молитвенную	
3.	и смысла тварного мира, а через него — к созерцанию вечности: вечной	
4.	собой человеческую душу и, переливаясь через край, исполняет, наполняет собой и	
5.	до краёв и снова переливается через край.	
6.	и дадим Ему просто любить через нас, действовать через нас, быть	
7.	Сигнал от перемещения ручки управления через блок запаздывания подавался на монитор	
8.	Через меридианы, где разность углов запаздывания	
9.	Траектория, проходящая через любую начальную точку, представляет полную	
10.	Через дефектную изоляцию может проходить ток	
11.	величиной является возвратное напряжение, измеренное через 30 с после отключения изоляции	
12.	Отфильтрованные через гидроциклон и джутовую ткань стоки	
13.	Повторно политуру наносят через 2…3 мин после нанесения	
14.	возможно, только 50% переводов осуществляется через официальные каналы (11).	

Рис. 7 Вставка наименования значения контекста

Установив в соответствующую ячейку курсор и щелкнув левой клавишей на нужном термине в буфере обмена, вставляем термин в ячейку. После того, как таблица заполнена, произвести подсчет значений. Для этого используется функция «выделение при чтении». Выделить столбец «значение». В поисковое окно вкладки найти вставить наименование значения. Система выделит искомые термины и сообщит число выделений (рис 8).

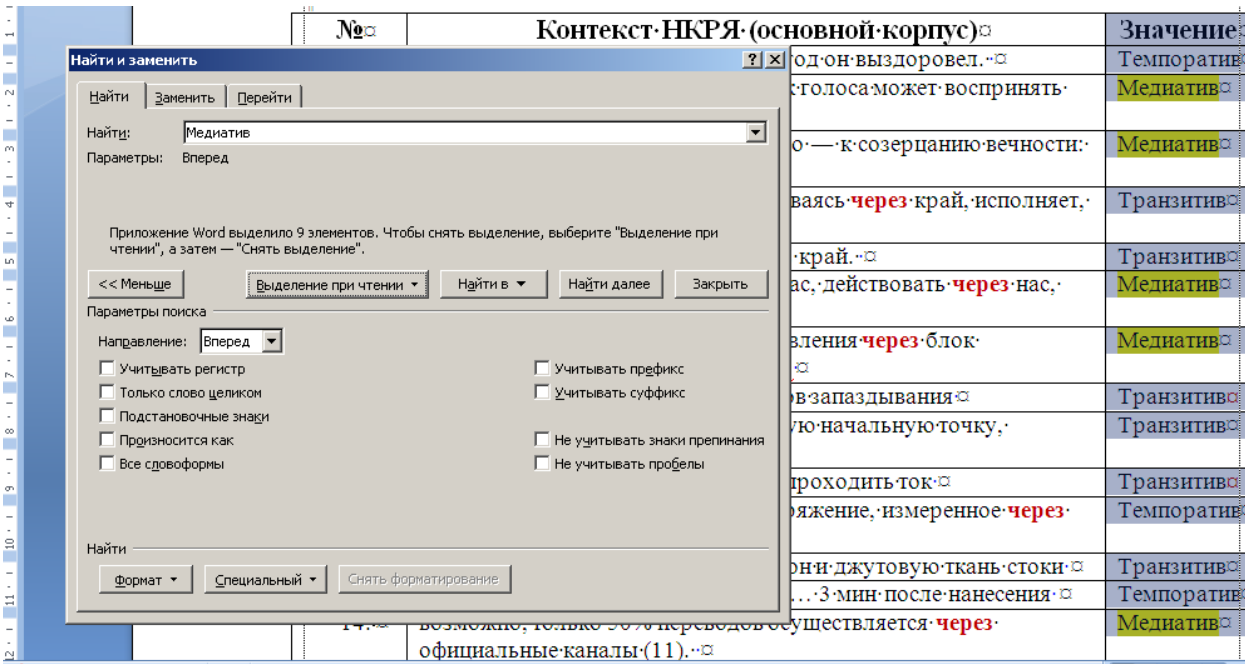


Рис. 8 Определение числа значений «медиатив»

Результаты подсчетов оформляются следующим образом

Основной корпус

Объем всего корпуса: 115 645 документов, 23 803 881 предложение, 283 431 966 слов.

через

Найдено 224 145 вхождений.

Отобрано 204 контекста

Значение	Число контекстов	% от общего числа контекстов
темпоратив	103	51,0
транзитив	61	30,4
медиатив	35	17,6
дименсив	4	2,0
фразеологизм	1	0,5

Сводная таблица значения предлога «через» по корпусам, % от числа отобранных

Корпус НКРЯ	темпоратив	транзитив	медиатив	дименсив	фразеологизм
Основной	51,0	30,4	17,6	2,0	0,5
Газетный	49,0	16,5	35,2	0	0,4
Устный	48,0	27,5	23,2	0,5	0,9
Поэтический	23,2	70,4	5,5	0	1,2