ОТЗЫВ

о выпускной квалификационной работе аспирантки Е.О. Косаревой

**«СТАТИСТИЧЕСКАЯ ДИНАМИКА ЛЕКСИЧЕСКИХ ЕДИНИЦ**

**В ПОВСЕДНЕВНОЙ РЕЧИ**

**(НА МАТЕРИАЛЕ КОРПУСА "ОДИН РЕЧЕВОЙ ДЕНЬ")»**,

Настоящая рецензия построена как комментарий к основным частям представленной к защите ВКР (введению, двум главам, заключению) и содержит последовательное изложение содержания работы с отзывами хвалебного и критического характера.

Целью исследования, по выражению автора, является «*описание динамики статистических показателей в лексике современной повседневной речи*» (С. 4). Описание выполняется на материале выборки из речевого корпуса. Наблюдение за динамикой производится при увеличении используемой для составления частотных списков выборки с 10 тыс. до 200 тыс. слов с шагом в 10 тыс. слов.

Во Введении автор ставит перед собой весьма масштабные задачи, в частности, планируя «*Проанализировать особенности повседневной речи*» (С. 4). Одна из задач – «*Охарактеризовать звуковой корпус 'Один речевой день'*». Не совсем понятно, почему характеристика корпуса, многажды прописанная (и не аспирантом) во множестве статей, фигурирует в числе его «задач»?

Теоретическая часть представляется относительно добротной, но несколько простоватой для работы аспирантского уровня. Так, для определения основных понятий статистической лексикографии автор обращается к двум **учебным пособиям**, это «Статистическая лексикография»1975 г. и «Частотные словари» 2001 г. П.М. Алексеева. Из «видов частотной характеристики» (С. 9) Екатерина Олеговна упоминает **только** об абсолютной частоте, о количестве вхождений на миллион слов коллекции, которое в работе названо «*частотой на миллион словоформ*» и об инвертированном индексе, который назван «*количеством источников, в котором данные единицы встречаются*». Отсутствует информация о традиционно используемых коэффициентах, например, коэффициенте правдоподобия (хотя, например, в том же частотнике НКРЯ представлен LL-score, см. «Словарь значимой лексики художественной литературы»), не упомянут, в частности, и коэффициент вариации D А. Жуйана.

В разделе 1.1.3 (это страницы 9 и далее) мне как читателю не хватило ссылок на работы О. Ляшевской и С. Шарова, предваряющих выпуск их частотного словаря. Публикации упомянутых авторов Екатерина Олеговна явно читала (что, безусловно, хорошо). В частности, на С. 10 видим «*частоты таких характерных для этой исторической эпохи слов как партия или товарищ в них* [в словарях Штейнфельдт и Засориной] *сопоставимы с частотами служебных слов*». Это – информация из статьи «Частотный словарь НКРЯ: концепция и технология создания» 2008 г., но никак не наблюдения автора ВКР.

В этом же разделе есть и фактические неточности. Так, на С. 11 сказано, что *«‘Частотный словарь современного русского языка’* ***основан на материале корпуса НКРЯ объемом более 600 млн словоупотреблений***». Это противоречит действительности: на самом деле «объём выборки, на которой строится большинство разделов словаря, составляет 92 млн. словоупотреблений».[[1]](#footnote-1) Кроме того, создатели Корпуса русского литературного языка, НЕ отождествляют язык драмы с разговорной речью, как можно думать из изложения автора (см. С. 3, С. 15, С. 12, ср.: «*Последний* [подкорпус драмы] *рассматривается в качестве письменно зафиксированной разговорной речи*»), а только указывают, что драма является «**некоторым приближением** к зафиксированной на письме разговорной речи» (http://www.narusco.ru/STAT004/, см. slovar.rar, Введение).

Вторая часть теоретической главы (раздел 1.2), озаглавленная «*Особенности повседневной разговорной речи*», в начале является по преимуществу грамотным изложением бытующих в русистике взглядов на природу и особенности разговорной речи. Хотелось бы заметить только, что Земская и Китайгородская всё-таки считали разговорную речь некодифицированной разновидностью литературного языка, а не «*самостоятельным явлением*» (см. С. 16). Хотя автор, видимо, подспудно присоединяется к школе О.Б. Сиротининой, озаглавив раздел 1.2.1 «*Разговорная речь* ***в системе функциональных стилей***».

В то же время к разделу 1.2.2. «*Инвентарь языковых средств повседневной речи*», производящему угнетающее впечатление, возникает целый ряд вопросов. Что имеется в виду под «*бóльшей подвижностью морфем*» на морфологическом уровне (С. 21)? Что позволило автору утверждать, что в разговорной речи присутствует «***специфическая парадигма имени***»? Неужели автор думает, что новые звательные формы типа *мам*, *пап* наблюдаются только в разговорной речи (они, как минимум, присутствуют в языке современной художественной литературы, в текстах электронной коммуникации)? Что позволило автору утверждать, что в разговорной речи НЕТ кратких прилагательных, НЕТ причастий и деепричастий (С. 21)?

Не вполне ясно, что значит утверждение автора о том, что на синтаксическом уровне в разговорной речи используются «***преимущественно семантические средства связи***» (там же)? Тут стоит заметить, что обычно черты аналитизма находят всё же в морфологии, а не в синтаксисе разговорной речи. Также хотелось бы получить пояснения по поводу фразы «*В отличие от письменной речи, в устной разговорной возможно свободное соединение, наложение, и использование специфических союзов*». Наложение (если речь идёт о наложении речи) – не синтаксическое явление; и что имеется в виду под «*свободным соединением*»? Действительно ли это «*соединение*» (вид связи?) невозможно в письменной речи?

Читателю сложно интерпретировать и высказывание, относимое автором к «лексическому уровню»: «*В системе разговорной речи меньше так называемых «пустых клеток», чем в кодифицированном литературном языке*» (С. 22). Идёт ли речь о заполнении лакун в неполных парадигмах и образовании форм типа *победю*? Если да, то вряд ли это явление имеет отношение к лексическому уровню. Неясно, почему жаргонизмы и профессионализмы на С. 23 приравнены к заимствованиям и названы «*инородными вкраплениями*» по отношению к разговорной речи.

Заключительный раздел теоретической главы 1.3. «*Корпус ‘Один речевой день’*» вполне добротен. Выводы к теоретической главе (С. 35) сформулированы довольно удачно.

Вторая, практическая глава «Разработка и анализ Частотника ОРД» начинается с описания методики составления частотника. Методика была бы ясной, если бы не следующее описание: «*был составлен* ***частотный список*** *из 200 000 словоупотреблений*», «*список был разбит на 20 «порций» по 10 000 словоупотреблений каждая*» (С. 36). Видимо, стоит понимать это так: на порции был разбит **не частотный список, а выборка** из 200000 словоформ, то есть было выделено20 подвыборок из 10000 слов каждая, которые затем объединялись.

Затем автор пишет о понятии слова (имея в виду обозначить, что является единицей составляемого словаря). Как «источники» авторского списка неоднословных лексических единиц заявлены одновременно словари Рогожниковой, Ефремовой, МАС, словарь неоднословных лексических единиц НКРЯ и список «составных слов», предлагаемый создателями Корпуса русского литературного языка.

Характеризуемый далее состав частотника ясен – в него просто включалось всё, что с некоторыми допущениями можно считать словом и что встретилось в транкрипте. Не стоило бы только включать в состав «ошибок и оговорок» (это 7 категория единиц) просторечные формы *евоный* и *ихний* (С. 41).

На С. 42 частицы *ка*, *де*, *то*, которые вопреки правилам орфографии пишутся в транскрипте не через дефис, а через пробел, почему-то названы «***дублетами***». Хотелось бы знать, что имел в виду автор.

Далее указано, что лемматизация «*производилась вручную*» (С. 43). Насколько мне известно, это не так: лемматизация производилась с помощью анализатора mystem, а вот снятие омонимии выполнялось вручную. **Зачем** вообще аспиранту-матлингвисту в 2016 году писать о том, что он выполнил лемматизацию 200 000 слов вручную? Неясность и с частеречной разметкой. Автор пишет, что «*была снята частеречная омонимия*», но не пишет, как была произведена разметка (С. 43).

Вторая часть второй главы (раздел 2.2. «*Статистические характеристики Частотника ОРД*») содержит информацию о параметрах оценки полученных данных по частотности лексических единиц в речевом корпусе. Этот раздел ВКР показался мне качественным (возможно, это связано с тем, что я ничего не понимаю в статистике).

Интересна информация о том, что объём частотника в леммах составляет 13200 единиц, что первая тысяча лемм «*покрывает около 82% всего объема*» (видимо, объёма выборки из 200 000 слов, С. 46), что наиболее встречаемое имя собственное имеет 267 ранг, и что максимальной частотой обладает местоимение *я*, что *я* и *не* при наращивании объёма выборки сохраняют 1 и 2 место в частотном списке. Не стоило бы только подтверждать данные о частотности лемм данными о частотности словоформ по корпусу (так как Т.Ю. Шерстинова в упомянутой здесь публикации всё же учитывала словоформы, а не леммы, см. С. 46 о *я*). Бессодержательным представляется пассаж «*При этом* [при наращивании объёма] *остальные слова ведут себя по-разному: одни постепенно переходят преимущественно на более высокие уровни; другие, напротив, на более низкие; у третьих наблюдаются неустойчивые колебания в разных направлениях*» (С. 47).

Обсчитанные в разделе 2.2 показатели средней частоты, величи́ны медианного ранга, рангового среднего, индекса Хирша, индекса концентрации и золотого сечения не подвергаются интерпретации(автор отсылает читателя к таблице 3, заметим, что в работе имеется две таблицы 3). Остаётся просто смотреть в найденную таблицу и верить цифрам, хотя хотелось бы понять, что, по мнению автора, приведённые цифры значат.

Интересна информация о доле слов с однократным вхождением по отношению ко всем словам выборки. Автор пишет, что число однократных единиц для частотника ОРД составляет 45% (С. 48). Судя по таблице, коэффициент разнообразия К=0,45 для выборки в 140, 150, 170, 180, 190 и 200 тыс. слов.

Наконец, заключающий практическую главу раздел 2.3 направлен на преобразование числовой информации из второй таблицы 3 в графики. В этом разделе автор приводит определение графика (С. 54), много внимания уделяет (слишком) подробному обзору методов выравнивания динамических рядов (С. 54-56). Затем выбирает один из методов («*метод взвешенных скользящих средних*»), получает эмпирические и сглаженные кривые, представляет читателю соответствующий ряд графиков. Автор снабжает их чрезвычайно лапидарными комментариями относительно конфигурации кривых (С. 60 и далее).

Выводы к практической главе вызывают двойственное впечатление: с одной стороны, третий вывод «*Анализ динамики статистических рядов показывает, что для частотного списка повседневной речи существует три параметра (Частота самого частотного слова (Fmax), Медианный ранг (Meᵣ) и Золотое сечение (Gᵣ)), которые не зависят от объема выборки*» представляется ценным и вытекающим из предыдущего изложения. Что касается второго вывода, то, видимо, если «*оправданность использования*» ряда девяти статистических параметров была подтверждена, то не Екатериной Олеговной и не в рамках данной ВКР. Первый вывод «*при построении частотника повседневной речи важно учитывать специфику лексического материала*» представляется тривиальным.

Заключение адекватно основному содержанию работы.

Вдобавок к высказанным, при чтении ВКР возникли следующие замечания и вопросы:

В тексте встречаются довольно неприятные для аспиранта фактические ошибки, например, на С. 3 читаем, что НКРЯ является «*крупнейшим корпусом русского языка на сегодняшний день*». Это неверно. Объём всех корпусов в составе НКРЯ – 600 млн слов. Корпус русских текстов ruTenTen 2011 насчитывает 18 млрд. токенов, Araneum Russicum Maius – 1216 млн. токенов, Google books Ngram Viewer (Корпус книг на русском языке) – более 67 млрд. словоупотреблений, а размеры ГИКРЯ к текущему лету 2016 г. достигли 19,8 млрд. словоупотреблений.

На С. 8-9 безо всяких ссылок представлена типология частотных словарей. Неясно, является ли Екатерина Олеговна её автором. На С. 18 без ссылок представлены особенности монолога, диалога и полилога. Если эти «особенности» были выделены в результате чтения научной литературы, на источники стоило бы сослаться. Сказанное относится и к перечислению «*черт*» разговорной речи со С. 19, а также к выделению компонентов коммуникативной ситуации со С. 20.

При всех вопросах и замечаниях заявленная цель работы представляется достигнутой. Думается, что работа соответствует требованиям, выдвигаемым к выпускным квалификационным работам аспирантов и заслуживает положительной оценки.

К.ф.н., доц. О.В. Блинова

Представленная работа отвечает требованиям, выдвигаемым к выпускным квалификационным работам аспирантов.

К.ф.н., доц. О.В. Блинова

30.05.2016

1. См.: Шаров С.А., Ляшевская О.Н., Введение к частотному словарю современного русского языка, URL: http://dict.ruslang.ru/freq.pdf. [↑](#footnote-ref-1)