Ягунова Е.В., Пивоварова Л.М. Экспериментально-вычислительные исследования художественной прозы Н.В. Гоголя // XLII Виноградовские чтения в МГУ «В.В. Виноградов о художественном тексте»: Материалы. – М., 2012

Е.В.Ягунова, Л.М.Пивоварова

Экспериментально-вычислительные исследования художественной прозы Н.В. Гоголя

1. Постановка проблемы. Цели, гипотезы, задачи

К сожалению, в современных лингвистических парадигмах творческое наследие В.В.Виноградова либо догматизируется, либо недооценивается. Последнее зачастую характеризует экспериментальную и прикладную парадигму лингвистических исследований.

Мы хотим показать, как решались нами – в экспериментальных и прикладных парадигмах – некоторые из глобальных теоретических задач, сформулированные в свое время В.В.Виноградовым. В первую очередь речь идет о постановке задачи изучения языка русской художественной литературы и индивидуального стиля (языка) писателя (прежде всего, статья В.В.Виноградова «О языке художественной литературы» (Виноградов 1959)). Что такое язык художественной литературы? И чем он отличается, например, от языка научных или новостных текстов?

Кроме того, современная коллокативистика, методы которой используются в данной работе, во многом заимствовала свой подход из работ В.В.Виноградова «Об основных типах фразеологических единиц в русском языке» (1947) и «Основные понятия русской фразеологии как лингвистической дисциплины» (1946)[1].

Целью работы является сопоставительное исследование структур художественного сюжетного текста – на примере произведений Н.В. Гоголя – в сравнении со структурами научного и новостного текстов. Данная работа входит в большое исследование зависимости структуры текста от функционального стиля (а также предметной области, жанрово-стилистических особенностей и т.д.) (см., напр., Ягунова, Пивоварова 2010а; Пивоварова, Ягунова 2010; Ягунова, Пивоварова 2011б). Художественный текст представляет собой, с одной стороны, наиболее сложный и проблематичный материал для такого рода исследования, с другой – многие интересные результаты могут быть получены именно в сопоставлении результатов исследования художественного vs. научного текста (или художественного vs. новостного текста). В ходе такого рода сопоставления может быть выделена «собственно информационная составляющая», так как научные (или новостные) тексты реализуют прежде всего информационную функцию.

В лингвистике текста часто говорят о различении синтаксических, семантических, а также информационных (смысловых) структур текста (см. обзор в Ягунова 2008). Граница между этими структурами нечеткая. Попробуем идти не от привычного разделения на уровни языка и речевой текст. В наших работах предлагается исследование структур текста, соотносимых либо со стилевыми характеристиками (функциональным стилем, стилем конкретного автора или, например, стилем новостного источника), либо с тематикой текста. Терминология в такого рода работах пока не сложилась; предлагаем ориентироваться на используемую в наших работах терминологию, ориентированную первоначально на работу с текстами научного и новостного (газетно-публицистического) функциональных стилей (см., напр., Ягунова, Пивоварова 2010а; Ягунова, Пивоварова 2010б; Ягунова, Пивоварова 2011б). Семантической структурой мы называем структуру, характеризующую прежде всего стилевые характеристики (предварительно – научных и новостных текстов), информационной структурой – характеризующую тематику, предметную область анализируемых текстов. Тем более, что на уровне методики

[1] См. в (Виноградов 1977)

экспериментального (автоматического) выделения эти структуры соответствуют разным статистическим мерам. Итак, общие предварительные гипотезы этого исследования состоят в следующем:

- формально определяемые (на основании статистических мер) семантическая и информационная структуры лучше всего различаются для информационно насыщенных политематических коллекций;
- для художественных произведений (циклов) такого рода структуры могут быть выделены таким же формальным образом, но эти структуры тесно взаимодействуют, образуют сложное взаимопереплетение – в отличие от информационно насыщенных научных и новостных текстов.

Теоретически семантическая структура должна в наибольшей степени соотноситься со стилем (характерном для писателя, цикла, произведения), а информационная структура – с содержанием произведения и/или цикла. Основа для формирования семантической структуры текста (цикла, коллекции): набор коллокаций, выделяемых с помощью меры t-score (максимальные значения меры); основа для формирования информационной структуры: во-первых, набор коллокаций, выделяемых с помощью меры MI (mutual information, коэффициент взаимной информации) (максимальные значения меры), во-вторых – ключевые слова, выделяемые в ходе вычислительного эксперимента (с помощью меры TF-IDF) и эксперимента с информантами (см. п.2).

Для монотематических – например, научных – коллекций (с соблюдением единого стиля за счет серьезной редакторской правки) и политематических коллекций возможны существенные различия: в первом случае коллокации, характеризующие стилевые и тематические характеристики, могут смешиваться.

Степень простоты и однозначности для процедур выделения ключевых слов зависит от следующих параметров:

– от функционального стиля текста (художественный, научный, новостной, официально-деловой),
– от темы, стиля, жанра и т.д.,
– от стиля конкретного писателя,
– от тематики произведения или цикла произведений рассматриваемого писателя
– от степени статичности vs. динамичности[2] повествования.

Почему коллокации, почему статистика? **Коллокации** понимаются нами как в значительной степени неслучайное сочетание двух и более лексических единиц, характерное для определенного текста (цикла, коллекции текстов). Традиционно выделяемые списки коллокаций отражают, главным образом, интуицию исследователя и лишь в некоторой степени могут быть соотносимы с изучением тех особенностей, которые не просто заложены в языке (всех текстах на этом языке), но в существенной степени зависят от типа рассматриваемых текстов. Альтернативой интуитивному методу можно считать использование различных статистических мер, позволяющих автоматически выделить из текстов коллокации и ранжировать их по степени устойчивости в соответствии со значениями выбираемых мер. Для нас статистический метод является единственно приемлемым, т.к. в нашем исследовании рассматриваются большие массивы текстов разных функциональных стилей и предметных областей, а список потенциальных коллокаций для них принципиально не задан, поскольку этот список является отражением тех языковых и экстралингвистических характеристик, которые заложены в анализируемых текстах., и выявление которых является конечной целью данного исследования.

Как мы понимаем структуры в данной работе? Под семантическими или информационными структурами понимаем распределение анализируемых коллокаций (топ-списков) на фоне всех прочих сочетаний слов текстов (цикла, коллекции), для

[2] Динамичности соответствует последовательность сменяющих друг друга ситуаций (напр., можно оценить количество ситуаций).

ключевых слов аналогично – распределение ключевых слов на фоне неключевых (всех прочих). Топ-списки определяются на основании анализа полученных выдач (коллокации или ключевые слова со значениями мер). В данной работе топ-списки коллокаций составляло около 100 единиц.

В исследованиях научных и новостных текстов были проверены и подтверждены следующие гипотезы (см. подробнее в Ягунова, Пивоварова 2010б; Пивоварова, Ягунова 2010; Ягунова, Пивоварова 2011б):

1. Используемые в работе статистические меры (MI и t-score) позволяют охарактеризовать предметную область и стилистические особенности новостных текстов;
2. Списки коллокаций, полученных с помощью MI и t-score, различны:
    a) коллокации, выделяемые с помощью MI, позволяют определять, прежде всего, наименования объектов, термины, сложные номинации, отражающие предметную область,
    b) критерий t-score направлен на выделение «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов) и «устойчивых конструкций», где и те, и другие характеризуют стилистические особенности новостных текстов;
3. Коллокации, выделяемые для монотематической коллекции (на примере научных текстов), характеризуются большей однородностью:
    a) коллокации, выделяемые с помощью MI, точно определяют предметную область, но могут включать и клише или стилевые маркеры (напр., *(на) наш взгляд, свою очередь, речь идет, представляет собой*);
    b) коллокации, выделяемые с помощью t-score дают представление о наборе общеязыковых устойчивых сочетаний (или, скорее, общих для рассматриваемой коллекции);
    c) степень тематической однородности коллекции соотносится с однородностью множества выделяемых коллокаций: терминологические коллокации, общие для всех (или подавляющего большинства) текстов коллекции, характеризуются высокими значениями меры t-score.

Третья гипотеза имеет для нас особое значение. Целью данной работы является изучение текстов Н.В.Гоголя, а не набора из любых (разных) текстов художественной литературы. Поэтому наши интересы лежат в области изучения близких к монотематичности текстов, циклов, коллекций (см. ниже три тематически наиболее однородные коллекции из произведений Н.В.Гоголя).

Еще одним интересующим нас параметром является степень статичности vs. динамичности повествования (нарратива), что отражено при отборе материала. Предполагается, что в рамках трех коллекций будет учтено разделение на потенциально более динамические и более статические. Анализируемые ранее научные и новостные тексты, очевидно, являются статическими. На уровне **дополнительной гипотезы**: 1) статические тексты (согласно свойству статичности) имеют семантическую и информационную структуры более близкие к структурам научных и новостных текстов; 2) динамические художественные тексты противопоставлены научным и новостным по двум параметрам: как художественные и как динамические.

Технические задачи, реализуемые на материале текстов Н.В.Гоголя[3]:

1. Выявление наиболее связанных коллокаций, характеризующих тематику текстов как элементов информационной структуры,
    - использование меры MI;
2. Выделение наиболее связанных коллокаций (клише и стилистические маркеры),

[3] В данной работе мы ограничимся описанием биграммных коллокаций (состоящих из двух слов) в силу заданного формата статьи.

характеризующих семантическую структуру,
    - использование меры t-score;
3. Выделение ключевых слов в ходе вычислительного эксперимента с использованием коэффициента важности tf-idf;
4. Выделение ключевых слов в ходе эксперимента с информантами.

Решение поставленных задач позволит сопоставить представление об информационной структуре текстов рассматриваемых коллекций, полученное в ходе решения задач 1, 3 и 4, и, далее, сравнить данные об информационной (задачи 1, 3, 4) и семантической (задача 2) структурах, сосуществующих и тесно переплетающихся в построении художественного текста у Гоголя.

2. Материал и методика

В качестве основного анализируемого материала рассматриваются 3 тематически наиболее однородные коллекции: 1) «Петербургский цикл», 2) «Мертвые души», 3) «Украинская тематика»: «Миргород» и «Вечера на хуторе близ Диканьки»[4].

В качестве материала для сравнения использовались три коллекции текстов (подробнее см. Пивоварова, Ягунова 2010):

- новостных: портала www.lenta.ru с апреля по декабрь 2009;
- научных:
    - материалов международной конференции «Диалог» «Компьютерная лингвистика и интеллектуальные технологии» за 2003-2009 годы;
    - материалов конференции «Корпусная лингвистика» 2004-2008 года (монотематическая коллекция).

Как уже было сказано, на данном этапе нами использовались две меры: для решения задачи 1 – MI (Church, Hanks 1990; Stubbs 1995), для решения задачи 2 – t-score (Church et al. 1991).

Мера MI является способом проверить независимость появления двух слов в тексте: если слова полностью независимы, то вероятность их совместного появления равна произведению вероятностей появления каждого из них, то есть произведению частот (использование абсолютных частот вместо относительных увеличивает значение MI для всех коллокаций в корпусе на константу, однако не меняет ее вероятностного смысла).

$$MI = log_2 \frac{f(c_1, c_2) \times N}{f(c_1) \times f(c_2)},$$

где

c_i – коллокаты;

$f(c_1,c_2)$ – абсолютная частота встречаемости коллокации c_1 c_2, с учетом порядка коллокатов внутри биграммы ;

$f(c_1)$, $f(c_2)$ – абсолютные частоты c_1 и c_2 в корпусе;

N – общее число словоупотреблений в корпусе.

Из определения видно, что мера MI зависит от размера корпуса: чем больше исследуемый корпус, тем выше в среднем получаемые по нему значения MI. Это свойство, видимо, должно отражать большую степень доверия к данным, полученным на материале большего корпуса. Однако в настоящем исследовании мера MI используется как средство ранжировать коллокации внутри одного корпуса по степени их связности – сравнение

[4] I. «Петербургские повести»: «Портрет», «Шинель», «Нос», «Невский проспект», «Коляска», «Записки сумасшедшего»; II. «Мертвые души»; III. Украинская тематика: «Вечера на хуторе близ Диканьки», и цикл «Миргород» («Вий», «Тарас Бульба», «Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем»).

между коллекциями осуществляется лишь по рангу, но не по значению меры для выделенных биграмм.

Другим недостатком меры MI, который отмечают многие исследователи (в том числе Stubbs 1995; Manning, Schutze 2002 и др.), является ее свойство завышать значимость редких словосочетаний, что делает данную меру совершенно «беззащитной» перед опечатками, иностранными словами и другим информационным шумом, который неизбежен в большой коллекции. Поэтому для данной меры используется порог отсечения, равный 16: в данной работе мы рассматривали только те биграммы, которые встретились в коллекции не менее 16 раз[5].

Необходимо отметить, что, как правило, при подсчете меры MI порядок слов внутри коллокации не учитывается – данная мера отражает взаимозависимость двух лексем и/или словоформ, но не значимость конкретной коллокации. В наших работах, однако, учитывался порядок коллокатов: мера MI подсчитывалась в отдельности для каждой конкретной пары лексем и/или словоформ.

Для решения задачи 2 нами использовалась мера t-score (см. об этой мере подробнее в (Church et al. 1991; Stubbs 1995)), которая учитывает частоту совместной встречаемости целевого слова и его коллоката. Она отвечает на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами. Мера t-score, расчитывается по формуле (условные обозначения здесь приняты те же, что и выше для MI):

$$t-score = \frac{f(c_1, c_2) - \frac{f(c_1) \times f(c_2)}{N}}{\sqrt{f(c_1, c_2)}}$$

Данная мера используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Значение данной меры тем выше, чем выше частота коллокации в коллекции. Данная мера содержит коррекционный компонент (вычитание деленного на размер коллекции произведения частот коллокатов), но эта поправка отражается лишь на самых частотных словах. Это свойство часто делает данную меру малопригодной для поиска терминологических словосочетаний и для этой цели она, как правило, не используется.

Для решения задачи 3 нами использовалась мера TF-IDF; это традиционная статистическая мера, применяемая для оценки важности слова в контексте документа, являющегося частью коллекции документов. Мера TF-IDF является произведением двух сомножителей: TF и IDF.

TF (*term frequency* — частота слова) оценивает важность слова t_i в пределах отдельного документа:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k},$$

где n_i есть число вхождений слова в документ, а в знаменателе — общее число слов в данном документе.

IDF (*inverse document frequency* — обратная частота документа) — инверсия частоты, показывающая количество документов коллекции, в которых встречается некоторое слово. Учёт IDF уменьшает вес широкоупотребительных слов (слов, встретившихся во многих документах коллекции):

[5] Для материалов «Корпусной лингвистики», также как и для произведений (циклов) Н.В. Гоголя, порог отсечения равен 16, для больших по объему коллекций портала Лента.ру и материалов конференции «Диалог» – порог равен 40. Порог подбирался эмпирически.

$$\mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|},$$

где |D| — количество документов в корпусе; $|(d_i \supset t_i)|$ — количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. На основании весов слов – значений меры – мы можем определить потенциально ключевые слова. Правильность этого определения зависит, главным образом, от того, насколько правильно определен контекст, то есть коллекция, с которой сравниваются слова интересующего нас текста или – как в случае данной работы – произведения максимально однородной подколлекции (цикла)[6].

Анализируемые подколлекции (циклы) текстов Н.В. Гоголя сопоставлялись с контекстом: коллекцией, включающей уже перечисленные тексты Н.В.Гоголя и сборники А.П.Чехова[7] «Человек в футляре», «Рассказы 1887 год», «Рассказы. Повести. 1888-1891», «Рассказы. Повести. 1892-1894», «Рассказы. Повести. 1894-1897». Состав контекста (выбор произведений А.П.Чехова, входящих в контрастивную коллекцию) обусловлен задачей получения максимально однородной контрастивной коллекции[8].

Для решения задачи 4 мы использовали традиционную методику проведения эксперимента с информантами со стандартной инструкцией А.С. Штерн (Мурзин, Штерн 1991): *Вспомните «Петербургские повести» Н.В.Гоголя. Подумайте над их содержанием. Выпишите 10-15 слов, наиболее важных для их содержания*. И далее также – «*Вспомните «Мертвые души» Н.В.Гоголя. ...*» , «*Вспомните украинский цикл Н.В.Гоголя («Миргород» и «Вечера на хуторе близ Диканьки»*)...». Единственное отличие от традиционного варианта заключалось в том, что информантам предлагалось вспомнить тексты, то есть оценивалось остаточное знание текста. В экспериментах по определению КС участвовало по 22 информанту для каждого из трех циклов. В качестве информантов выступали профессиональные филологи (не студенты), хорошо знающие русскую классику. К участию в эксперименте не привлекались преподаватели русской литературы в школе или ВУЗе, чтобы образовательные методики, программы, стандарты не влияли на результат эксперимента[9].

3. Результаты. Обсуждение результатов

В таблице 1 приводится топ-список (коллокации с максимальным значением меры MI), являющийся пересечением топ-списка для словоформных и лексемных биграмм (отдельно для «Петербургских повестей», «Мертвых душ» и «Украинской тематики»). Подробнее о причинах выбора тех коллокаций, которые выделяеются и для словоформ, и для биграмм см. в (Ягунова, Пивоварова 2010; Ягунова, Пивоварова 2011б).

В этой таблице (табл. 1а,б,в) представлены:

1. MI-коллокации, характеризующие тематику текстов, как элементов информационной структуры: напр., *Невского проспекта, коллежского асессора, статского советника, Акакию Акакиевичу*, *Павел Иванович, Миргородского повета, Хома Брут*;
2. MI-коллокации, соотносимые с составными и дискурсивными словами, клише,

[6] Коллекция как контекст для определения весов конкретных слов иногда называется контрастивной коллекцией, то есть текст (цикл, коллекция), для которой определяются веса слов, является фигурой, а коллекция, служащая контекстом, выступает в качестве фона (в терминах гештальт-психологии).

[7] Источник: А.П. Чехов. Полное собрание сочинений и писем в 30-ти томах. Сочинения. Том 1. М., "Наука", 1983

[8] Предварительный анализ позволил выбрать произведения А.П.Чехова и последующий анализ результатов с разными контрастивными коллекциями подтвердил правильность этого выбора.

[9] Решение задач выделения ключевых слов по данным методикам (вычислительного эксперимента и эксперимента с информантами) было отработано на материале научных и художественных текстов (см. Ягунова 2010а, Ягунова 2010б).

стилистическими маркерами: напр., *большей частию* (компонент[10]), *крайней мере, никоим образом, понимаете ли, таким образом, Боже мой, очень приятно, слава Богу, …нас черненькими…* (как компонент крылатой фразы);

3. MI-коллокации, представляющие собой предикативные конструкции: напр., *частию лежал, сказал, вы встретите, носит царица*.

Между названными тремя классами существуют пересечения и неоднозначность интерпретации. Напр., *понимаете ли* – это вводная и предикативная конструкция, часто характеризующая особенности стиля того или иного автора. Первый тип MI-коллокаций максимально соответствует тому типу единиц – прежде всего, сложных номинаций – который был выявлен на разных научных и новостных коллекциях (в рамках исследования функционального стиля).

Для сравнения приведем топ-списки MI-коллокаций (в порядке убывания меры):

- для коллекции новостных текстов портала Лента.ру за 2009 год: *Бритни Спирс, Эльвира Набиуллина, Ле Бурже, Лионель Месси, мысе Канаверал, бин Ладена, Норильского никеля, дельты Нигера, Ак Барс, тротиловом эквиваленте, тройскую унцию, Ролан Гаррос, дель Торо, дель Потро, Арбат Престиж, РАО ЕЭС, Салават Юлаев, Арсений Яценюк, голубых фишек, адронного коллайдера*;
- для научных текстов:
    - политематической коллекции материалов конференции Диалог за 2003-2009 годы: *ударном слоге, концептуальных графов, внешним посессором, оперативной памяти, вокального жеста, крайней мере, XIX века, лингвистического процессора, положение дел, первую очередь, картине мира, множественного числа, интеллектуальные технологии, корпусная лингвистика, отглагольных существительных, знаки препинания, педагогической коммуникации, основного тона, машинного перевода, устойчивых словосочетаний;*
    - монотематической коллекции из материалов конференции «Корпусная лингвистика» за 2004, 2006, 2008 годы: ***наш взгляд, (по) крайней мере**, речевой деятельности, художественной литературы, первую очередь, общим объемом, корпусная лингвистика, имена собственные, математической лингвистики, словарной статьи, **свою очередь**, предметной области, машинного перевода, точки зрения, **за счет**, **речь идет**, **прежде всего**, большое количество, настоящее время, **представляет собой**, млн словоупотреблений, **другой стороны**, семантических состояний, **одной стороны, таким образом**, разрешения неоднозначности, английский язык, **кроме того**, Национальный корпус, грамматических категорий, устная речь, база данных, **во многих**, лексических единиц, **дает возможность, зависит от, отличие от**, русский язык, корпусные данные, **отличается от, зависимости от, работы над**, частей речи, **во всех, при помощи**, морфологической разметки*.

Полужирным шрифтом – для новостных и научных топ-списков MI-коллокаций – выделены те коллокации, которые находятся на пересечении того, что относится к информационной структуре, и того, что относится и к семантической структуре (сочетаемостно выделяемые для этих коллекций – разной степени не только тематической, но и стилевой однородности – составные и дискурсивные слова, клише, близкие к ним устойчивые единицы).

В таблицах 1 и 2 полужирным шрифтом выделены те единицы, которые мы интерпретируем как пересечение семантической и информационной структуры; к ним примыкают предикативные единицы, выделенные курсивом, их интерпретируем как потенциальное пересечение семантической и информационной структуры (глагольные конструкции редко попадают в определение тематики текста).

---

[10] Большие конструкции разбиваются на биграммы, напр., *большей частию лежал на кровати – большей частию, частию лежал.*

Результаты сравнения позволяют делать выводы о многих факторов, влияющих на возможность разделения информационной и семантической структур текстов разных стилей. Включение в набор анализируемого материала монотематической коллекции научных текстов после редакторской правки позволяет говорить о том, что в определенных случаях и по определенным параметрам монотематические научные и художественные коллекции обнаруживают схожие свойства.

Таблица 1. *Топ-список MI-коллокаций*[11]

а. Петербургские повести

[11] Для простоты восприятия в таблице представлены словоформные биграммы, упорядоченные по убыванию значения меры.

пп	MI-биграмма
1	Невского проспекта
2	коллежского асессора
3	статского советника
4	12 часов
4	господами офицерами
5	Петербургские повести
6	ma chere
7	Акакию Акакиевичу
9	Милостивый государь
10	**большею частью**
11	милостивый государь
12	титулярный советник
13	начальник отделения
14	*частью лежал*
15	коллежский асессор
20	умоляющим голосом
22	передо мною

пп	MI-биграмма
24	**крайней мере**
26	ваше превосходительство
27	маиора Ковалева
28	значительного лица
29	сорок копеек
30	друг Гофман
31	**сих пор**
33	Иван Яковлевич
34	Андрей Петрович
37	поручик Пирогов
41	**такой степени**
44	**Отец мой**
47	**таким образом**
50	**каждый день**
51	**никаким образом**
54	перед зеркалом
55	мой друг

пп	MI-биграмма
57	новая шинель
59	рублей сорок
61	во внутрь
62	после обеда
65	**без сомнения**
66	**во сне**
67	**Боже мой**
69	**между тем**
70	***может быть***
73	***понимаете ли***
74	**однако ж**
76	**несколько минут**
77	**вам угодно**
80	два года
81	молодой человек
82	вместо носа
87	**про себя**

б. Мертвые души

пп	MI-биграмма
1	Кифа Мокиевич
2	Мокий Кифович
4	*ездят холостяки*
5	земская полиция
6	Павел Иванович
7	воскресным дням
8	врачебной управы
9	полковнику Кошкареву
10	Александра Степановна
11	Настасья Петровна
12	действительный статский
13	Софья Ивановна
14	Фома Большой
15	Фома Меньшой
16	губернаторскую дочку
17	ранней редакции
18	*увезти губернаторскую*
19	окончание главы
20	красного дерева
21	наваринского пламени

пп	MI-биграмма
22	трактирного слуги
23	второго тома
24	**слава Богу**
25	первом издании
26	близкий приятель
28	генерала Бетрищева
29	издании второго
30	французский язык
31	русскому обычаю
32	карточная игра
33	немецкого писателя
34	фраке наваринского
35	дядя Митяй
36	статский советник
37	Анна Григорьевна
38	Платон Михалыч
42	записной книжки
43	расположении духа
44	Андрей Иванович
46	двенадцатого года

пп	MI-биграмма
47	Константин Федорович
48	председателя палаты
49	Брат Василий
50	хозяйственная часть
51	мертвые души
53	книжки Н
54	полковник Кошкарев
55	*рукописи отсутствуют*
56	среди волн
57	капитан Копейкин
59	Афанасий Васильевич
60	ваше сиятельство
61	десять миллионов
62	Петр Петрович
63	Александр Петрович
64	**некотором роде**
65	Иван Антонович
66	Иван Григорьевич
70	**крайней мере**

в. Украинская тема

пп	MI-биграмма
1	Миргородского повета
2	Хавронья Никифоровна
3	глиняная кружка
4	смертным часом
5	ученики старших
6	Антона Прокофьевича
7	Демьян Демьянович
9	учеников младших
10	Агафия Федосеевна
12	Степана Кузьмича
13	большею частию

пп	MI-биграмма
15	младших классов
16	любезные читатели
17	Тиберий Горобець
18	изо рта
20	*носит царица*
21	Василиса Кашпоровна
22	Мосий Шило
23	гоп трала
24	старших классов
25	блаженной памяти
26	вывороченном тулупе

пп	MI-биграмма
27	клок волос
29	Григория Григорьевича
30	Фомы Григорьевича
33	милостивый государь
34	пшеничной муки
35	гусиный хлев
36	село Хортыще
37	Черное море
38	длинный клок
45	*понюхать табаку*
47	городские ворота

пп	MI-биграмма
49	рюмку водки
50	**крайней мере**
51	тысячи червонных
57	куренной атаман
62	Хома Брут
63	**той поры**

пп	MI-биграмма
64	есаул Горобець
67	*разинул рот*
70	Иванов сын
71	Никифоров сын
81	собачий сын
83	вороном коне

пп	MI-биграмма
90	Катеринин отец
101	об одолжении
104	*выступил вперед*
106	**Кой черт**

В таблице 2 приводится топ-список (коллокации с максимальным значением меры t-score), являющийся пересечением топ-списка для словоформных и лексемных биграмм (отдельно для «Петербургских повестей», «Мертвых душ» и «Украинской тематики»). В таблице отдельно приводятся значения частотности (частоты встречаемости (fr)) и меры t-score (t). В отдельных случаях поправочный коэффициент t-score корректирует значения частотности.

Для сравнения приведем топ-списки t-score-коллокаций (в порядке убывания меры):

- для коллекции новостных текстов портала Лента.ру за 2009 год: *об этом, по словам, а также, со ссылкой, ссылкой на, по данным, кроме того,* ***РИА Новости****, этом сообщает, при этом, в том, в России, во время, пока не, о том, в результате, настоящее время, миллионов долларов, связи с, сообщает РИА, в результате, в частности, миллиарда долларов, как сообщает*;
- для научных текстов:
  - политематической коллекции материалов конференции Диалог за 2003-2009 годы: *и т. (д.), может быть,* ***русского языка****, а также, в том, так и, на основе, и др,* ***русском языке****, таким образом, не только, в качестве, с помощью, в русском, могут быть, в виде, при этом, точки зрения, но и, в тексте, в частности, то есть, при этом, в рамках, о том, и не, в этом, а не, в данном, кроме того, в которых, и их, как в, в случае, а в, как и, из них, отличие от, и его, представляет собой, не может,* ***предметной области****, с точки, так как, только в, в качестве, зависимости от, в результате, этом случае*;
  - монотематической коллекции из материалов конференции «Корпусная лингвистика» за 2004, 2006, 2008 годы: *и т. (д)., может быть, а также,* ***русского языка****, в том, в корпусе, и в, так и, не только, таким образом, и др, точки зрения, на основе, могут быть, в тексте, настоящее время, в качестве, в виде, в рамках, том числе,* ***корпуса текстов****, в частности, с помощью, в словаре, при этом, с точки, при этом, и для, прежде всего, в текстах, в этом, кроме того, представляет собой, текстов в, слов в, слова в, так как,* ***английского языка****, соответствии с, в контексте, как в,* ***машинного перевода****, как правило, связи с, то же, а не, и пр, только в,* ***части речи****, в котором, не менее, слов и, текстов и, в настоящее, в которых,* ***параллельных текстов****, с использованием, в настоящее, в целом, из них,* ***корпус текстов****, именно, в соответствии, при создании, первую очередь,* ***предметной области****, в случае, другой стороны,* ***лексических единиц****.*

Полужирным шрифтом – для новостных и научных топ-списков t-score-коллокаций – выделены те коллокации, которые находятся на пересечении того, что относится к семантической структуре, и того, что относится и к информационной структуре (частотные для этих коллекций – разной степени монотематизации – неоднословные термины).

Таблица 2. *Топ-список t-score-коллокаций*

а. Петербургские повести

t-score биграмма	fr	t
как будто	61	7,72
Акакий Акакиевич	53	7,27
потому что	53	7,08
не мог	42	6,26

t-score биграмма	fr	t
может быть	37	6,06
будто бы	36	5,94
не было	44	5,87
Иван Яковлевич	28	5,29

t-score биграмма	fr	t
не могу	26	4,97
ваше превосходительство	24	4,90
вместе с	24	4,80

t-score биграмма	fr	t
Акакия Акакиевича	23	4,79
никак не	24	4,77
так что	28	4,67
это время	22	4,61
если бы	22	4,61
однако же	21	4,56

t-score биграмма	fr	t
никогда не	22	4,55
сказал он	24	4,49
вовсе не	22	4,47
Невский проспект	19	4,36
так же	20	4,27
крайней мере	18	4,24

t-score биграмма	fr	t
тут же	18	4,22
всё это	19	4,21
можно было	18	4,17
не может	20	4,16

б. Мертвые души

t-score биграммы	fr	t
сказал Чичиков	116	10,5
потому что	105	10,0
как бы	97	9,22
что ж	85	6,68
тут же	82	8,96
Павел Иванович	77	8,77
может быть	75	8,64
если бы	62	7,72
ничего не	64	7,59
как будто	59	7,57
самом деле	56	7,48
однако же	55	7,34

t-score биграммы	fr	t
в самом	55	7,19
не мог	54	7,09
никак не	52	7,01
однако ж	49	6,97
все это	53	6,92
так что	64	6,74
то есть	46	6,68
между тем	44	6,61
ваше превосходительство	40	6,32
про себя	38	6,15
еще не	49	6,08
это время	38	6,03

t-score биграммы	fr	t
вместе с	34	5,72
тот же	33	5,67
в городе	34	5,67
крайней мере	32	5,66
таким образом	31	5,56
по крайней	29	5,36
то же	33	5,34
Афанасий Васильевич	28	5,29
мертвые души	28	5,29
так сказать	29	5,24
несмотря на	25	4,93

в. Украина

t-score биграммы	fr	t
Иван Иванович	183	13,5
как будто	128	11,15
Иван Никифорович	147	9,97
потому что	91	9,37
Иван Федорович	70	8,34
вместе с	62	7,73
на свете	61	7,60
ничего не	61	7,53
что ж	76	7,37
не мог	56	7,28
однако ж	81	7,06

t-score биграммы	fr	t
может быть	50	7,06
если бы	50	6,96
это время	49	6,92
никто не	44	6,43
тут же	42	6,41
про себя	39	6,22
да и	55	6,14
несмотря на	39	6,13
самом деле	37	6,08
пан Данило	37	6,07

t-score биграммы	fr	t
Ивана Ивановича	36	5,99
между тем	34	5,81
можно было	34	5,76
на землю	34	5,58
в самом	33	5,54
со всех	31	5,54
еще не	44	5,41
всех сторон	29	5,38
никогда не	31	5,37

В таблице 3 приведены потенциально ключевые слова, выделенные с использованием коэффициента важности TF-IDF, слова упорядочены по убыванию значения этой меры. Пороговое значение определялось эмпирически.

В общем и целом, можно сказать, что определяемые таким образом слова представляют собой наименования действующих лиц, мест и событий. Полужирным шрифтом выделены слова, относящиеся к пересечению множеств ключевых слов, выделяемых в ходе вычислительного эксперимента (см. табл. 3) и в ходе эксперимента с информантами (табл. 4).

Для научных текстов предлагаемая методика дает еще более четкие результаты выделения и классификации ключевых слов (Ягунова 2010а; Пивоварова, Ягунова 2011а). Различие между художественными и научными текстами состоит, прежде всего, в весах этих признаков. В частности, различительная сила слова, оцениваемая с использованием третьего формального признака (TF-IDF), гораздо выше для научного текста, чем для художественного.

Таблица 3. *Ключевые слова, полученные в результате вычислительного эксперимента*

Питерские повести	Мертвые души	Украинская тематика
Акакиевич	**Чичиков**	**козак**
Ковалев	**Ноздрев**	**Никифорович**
Акакий	**Манилов**	пан
Яковлевич	Селифан	хата
маиор	**Собакевич**	запорожец
Шиллер	Костанжогло	**козацкий**
Чартков	человек	Андрий
Пискарев	**Плюшкин**	Тарас
проспект	Платон	Остап
Чертокуцкий	Хлобуев	Данило
чорт	тентетник	курень
портрет	слово	Катерина
человек	рука	**Иван**
Невский	тентетников	Бульба
глаза	время	парубок
Гофман	Копейкин	**Днепр**
рука	Мураз	дьяк
лицо	Антонович	**черевички**
медж	Петрушка	рука
пуф	**бричка**	**Чуб**
нос	Платонов	**кузнец**
квартальный	лицо	свитка
бакенбарды	купчая	**Голова**
время	Павел	галушка

департамент	город	Левко
голова	сторона	Оксана
комната	глаз	Янкель
художник	Кошкарев	хлопец
слово	место	Петро
Испания	ассигнация	сотник
штаб-офицерша	герой	человек
беспрестанный	несколько	лях
шинель	**души**	**Вакула**
ростовщик	дама	Миргород
асессор	голова	**Солоха**
коллежский	Леницын	Хома
титулярный	поэма	есаул
коломна	чубарый	**панночка**
лорнет	думать	Григориевич
прыщик	Иванович	куренной
Рафаэль	жизнь	Прокофиевич
Фидель	Бог	гетьман
Психея	дом	Дорош
происшествие	барин	комиссар
чиновник	полицеймейстер	**Иванович**
дама	председатель	шинок
казаться		

В таблице 4 приведены результаты эксперимента с 21 информантом по выделению ключевых слов, количественные данные приведены в абсолютных числах (указывается число информантов, записавших в анкете данное слово с точностью до лексемы).

Таблица 4. *Ключевые слова, полученные в результате эксперимента с информантами*

Питерские повести		Мертвые души		Украинская тематика		Украинская тематика (продолж.)	
слова	КС	слова	КС	слова	КС	слова	КС
шинель	14	помещик	5	черт	8	Вий	3
нос	8	дорога	8	ночь	8	**Днепр**	3
художник	10	тройка	8	**панночка**	6	еда	3
чиновник	10	**бричка**	7	**черевички**	6	звезды	3
Невский	9	Коробочка	7	**кузнец**	6	**казак**	1
Акакий	8	**Плюшкин**	7	любовь	5	нечисть	3
проспект	7	**Чичиков**	7	Рождество	5	парубок	2
сумасшествие	7	купчая	6	гусак	5	русалка	2
портрет	5	**Манилов**	6	ведьма	3	смех	3
Петербург	6	**Собакевич**	6	**Голова**	4	Украина	3
мечта	4	мертвые	5	**Иван Иванович**	4	хутор	3
майор	5	**Ноздрев**	5	**Иван Никифорович**	3		
страх	4	крепостные	3	Ивана Купала	2		
холод	4	Россия	3	праздник	4		
Акакиевич	3	губернатор	2	**Солоха**	4		
обман	2	**души**	2	**Чуб**	4		
Пирогов	3			ярмарка	4		

Пискарев	3			**Вакула**	3		

Наибольший интерес представляют слова (выделенные п/ж шрифтом), относящиеся к пересечению множеств ключевых слов, определяемых в ходе вычислительного эксперимента (см. табл. 3) и в ходе эксперимента с информантами (табл. 4).

Для вычислительного эксперимента имеют существенное значение такие факторы, как частотность слова в тексте, число документов, содержащих это слово, даже наличие/отсутствие очевидной внутренней формы (напр., Коробочка).

Слова, являющиеся «символами текста», далеко не всегда могут определяться в ходе вычислительного эксперимента. Например, лексема «тройка» (в частности, «*Эх, тройка! птица тройка, кто тебя выдумал? знать, у бойкого народа ты могла только родиться…*») встречается 13 раз в тексте (низкое значение компонента TF); однако, вряд ли кто-нибудь усомнится в значимости этого ключевого слова для нашего представления о тексте «Мертвые души» (8 человек из 21 записало это слово в своей анкете). Слово «дорога» является частотным в русском языке и, в частности, в текстах Н.В. Гоголя и А.П.Чехова. В «Мертвых душах» эта лексема встречается 119 раз, но оно встречается в большом количестве документов анализируемой коллекции, и за счет компонента IDF слово не попадает в ключевые. По мнению же информантов слово является ключевым (опять же 8 человек из 21 его записало в анкетах).

4. Заключение

В статье представлены результаты анализа семантической и информационной структур, где первая в наибольшей степени соотносится со стилем (характерном для писателя, цикла, произведения), а вторая – с содержанием произведения и/или цикла. Объекты исследования: цикл «Петербургские повести», поэма «Мертвые души» и произведения украинской тематики (циклы «Миргород» и «Вечера на хуторе близ Диканьки»). Методика исследования: вычислительный эксперимент и эксперимент с информантами. Семантическая и информационная структуры анализировались через сопоставление наборов коллокаций (двух типов, выделяемых на основании статистических мер MI vs. t-score) и ключевых слов.

В художественном тексте в результате взаимодействия и пересечения семантической и информационной структур семантическая структура приобретает элементы, свойственные содержательной стороне (например, частотные ключевые слова или коллокации становятся также характеристикой стиля), а информационная структура начинает включать те стилевые сочетания, которые приобрели важную для содержания роль. Такого рода взаимопроникновение отличает художественную прозу от информационно насыщенных стилей текста (научного и новостного). Проведенное исследование (сопоставление разных списков словосочетаний и слов) позволяет формальным образом охарактеризовать особенности построения анализируемых произведений Н. В. Гоголя.

Кроме общего противопоставления язык художественной литературы vs. научных текстов vs. новостных текстов мы рассмотрели дополнительные параметры. Существенную роль на взаимодействие семантической и информационной структур оказывают 1) степень тематической и стилевой однородности и 2) степень статичности/динамичности текстов. Так, в статье были кратко показаны основные различия во взаимодействии семантических и информационных структур в зависимости от выбора одной из трех коллекций («Петербургские повести», «Мертвые души» и «украинская тематика»), где противопоставление происходит как по степени однородности, так и по степени статичности/динамичности.

«Петербургские повести» отличаются взаимопроникновением структур, а списки потенциально ключевых слов, выделяемых на основании вычислительного эксперимента и эксперимента с информантами (см. табл.1а и табл.2а), хорошо демонстрируют различия

между двумя типами информационных структур: извлекаемой человеком в процессе понимании текстов vs. автоматом при реализации процедур информационного поиска. Структуры подколлекции «украинская тематика» характеризуется максимальной неоднородностью. Данные, полученные на материале поэмы «Мертвые души», оказываются промежуточными между этими подколлекциями.

Литература

1. Виноградов В.В. О языке художественной литературы. - М., 1959. С. 84—166
2. Виноградов В. В. Избранные труды. Лексикология и лексикография. - М., 1977
3. Мурзин Л. Н., Штерн А. С. Текст и его восприятие.– Свердловск, 1991.
4. Пивоварова Л.М., Ягунова Е. В. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) // *Материалы Симпозиума "Терминология и знание"* (Москва, май 2010 г.). М. 2010
5. Ягунова Е.В. *Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей).* Пермь. 2008
6. Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов // Материалы IV Международного конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». М., 2010а. – С. 533-534
7. Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика / Отв. ред. В.Т. Юнгблюд. Вып. 1. – Пермь, 2010б. С. 85-91.
8. Ягунова Е.В. Ключевые слова в исследовании текстов Н.В. Гоголя // Проблемы социо- и психолингвистики. Пермь, 2011. Вып. 15. Пермь 2011 (в печати)
9. Ягунова Е.В., Пивоварова Л.М. 2010а. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // *Научно-техническая информация, Сер.2, №6.* М. с.30-40
10. Ягунова Е.В., Пивоварова Л.М. 2010б. Извлечение и классификация коллокаций на материале научных текстов. предварительные наблюдения // *V Международная научно-практическая конференция "Прикладная лингвистика в науке и образовании" памяти Р.Г. Пиотровского (1922-2009) : Материалы.* СПб. С. 356-364
11. Church K., Hanks, P. 1990, ‘Word association norms, mutual information, and lexicogra-phy’, Computational Linguistics, 16(1), 22–29.
12. Church, K., W. Gale, P. Hanks and D. Hindle 1991 Using statistics in lexical analysis. In U. Zernik ed Lexical Acquisition. Englewood Cliff, NJ: Erlbaum. 115-64.
13. Manning C., Schutze H. Collocations // Manning C., Schutze H. *Foundations of Statictical Natural Language Processing*, 2002, pp.151-189
14. Stubbs M. Collocations and semantic profiles: on the case of the trouble with quantitative studies. // *Functions of language 2:11*, 23-55, Benjamins, 1995.