Применение Многомерного анализа к изучению языковой вариативности в русскоязычных Интернет-жанрах.

Катинская А. Ю.

Российский государственный гуманитарный университет

a.katinsky@gmail.com

Данная статья представляет исследование, находящееся в рамках других работ по автоматической идентификации Интернет-жанров и сосредоточенное на изучении взаимосвязи между языковыми признаками текстов и их жанрами на материале русского языка. С этой целью к небольшому, но многообразному с жанровой точки зрения корпусу Интернет-текстов на русском языке был применен Многомерный анализ, разработанный Дугласом Байбером в 1988 для изучения вариативности лингвистических признаков в регистрах английского языка (Biber, 1988). Данный анализ никогда не применялся для исследования языковых признаков русского языка. Мы обнаружили семь измерений вариативности языковых признаков и интерпретировали их с функциональной точки зрения. Также мы изучили распределение текстов в рамках девяти жанровых кластеров в пространстве данных измерений языковой вариативности. Результаты исследования показали, что каждый из представленных в корпусе жанров имеет индивидуальную характеристику в пространстве полученных измерений вариативности, что может применяться с целью лучшего понимания жанровой структуры русскоязычного Интернета, а также в целях автоматической жанровой классификации. Значения стандартного отклонения от средних значений оценок текстов по каждому из измерений могут быть использованы для определения ошибок в аннотировании, выполненном разметчиками, или для выявления в составе рассматриваемого жанра новых поджанров.

Ключевые слова: русский язык, Многомерный анализ, вариативность языковых признаков, жанровая классификация, функциональные размерности.

Multi-dimensional Analysis Applied to Study Linguistic Variation in Russian Web Genres

The paper presents a study continuing a line of works on automatic Web genre identification and focused on an investigation of relations between linguistic features and genres for Russian texts. For this purpose we applied the Multi-dimensional approach, developed by Biber in 1988 for studying register variation in English (Biber, 1988), to a small but genre diverse corpus of modern texts from the Russian Web. Multi-dimensional analysis has not been applied to studying linguistic features of Russian language before. We have discovered seven dimensions of linguistic feature variation and interpreted them functionally. We also studied the distribution of texts grouped into 9 genre classes in the space of these dimensions. The results showed that every genre in the corpus has its position in the multidimensional space of linguistic features that can be used for the purpose of better understanding of the genre structure of Russian Web and could be tested for the automatic genre classification. Deviations in genre dimension values for each text allow us to find errors in genre annotation and to detect fine-grained dimensional differences between subgenres.

Keywords: the Russian language, Multi-dimensional analysis, dimensions of register variation, genre classification, Functional text dimensions.

1. Введение

Автоматическое определение жанровой принадлежности текста является одной из задач, которую ученые пытаются решить на протяжении более двадцати лет, при том обстоятельстве, что и к самому определению жанра, и к пониманию системы жанров не выработано единого подхода. Так, некоторые исследователи создают практически необозримые классификации жанров, выделяя до нескольких тысяч категорий (Adamzik, 1995). В такой относительно новой и динамичной сфере речевого общения, как Интернет, сформулировать устойчивую систему жанров становится еще сложнее. Некоторые попытки исследовать жанры англоязычного Интернета были предприняты целым рядом ученых (Santini et al., 2010), для русского языка это проблема изучена в значительно меньшей мере.

В данной работе мы опираемся на нетрадиционный подход к жанровой классификации. Вместо закрытого набора жанров мы используем жанровые категории, выделяемые апостериорно на основе функционального сходства текстов между собой. Установление этого сходства осуществляется с помощью системы функциональных текстовых размерностей (Functional Text Dimensions, или FTD), разработанных С. А. Шаровым (Forsyth and Sharoff, 2014).

Мы принимаем гипотезу, согласно которой языковые признаки обладают разной частотой в текстах разных жанров и имеют тенденцию появляться в них группами, объединенными на основании взаимосвязи их коммуникативных функций. Все признаки в составе этих групп объединяются вокруг одной общей коммуникативной функции более высокого уровня, которую они совместно выполняют. Мы полагаем, что одним из оснований выделения любого речевого жанра являются данные функции. Следующим образом функциональный подход к определению жанров формулировал М. М. Бахтин в (Бахтин, 1986): «Определенная функция (научная, техническая, публицистическая, деловая, бытовая) и определенные, специфические для каждой сферы условия речевого общения порождают определенные жанры, то есть определенные, относительно устойчивые тематические, композиционные и стилистические типы высказываний». Таким образом, выбор жанровой формы высказывания определяется сферой речевого общения и задачами, которые говорящий ставит перед собой, то есть теми функциями, которые жанр должен реализовать. Мы полагаем, что жанровые функции, соответствующие коммуникативным потребностям говорящего, и вышеупомянутые общие функции, вокруг которых группируются различные языковые признаки, суть одно и то же. Нашей целью является попытка установить хотя бы подмножество данных общих функций, лежащих в основе ограниченного набора языковых признаков. С этой целью был реализован Многомерный анализ Д. Байбера, ранее не применявшийся к текстам на русском языке (и, по-видимому, к другим славянским языкам).

В разделе 2 данной статьи мы описываем методологию Многомерного анализа. Раздел 3 содержит описание корпуса, на котором было проведено жанровое исследование, а также основные принципы отбора языковых признаков и описание программного инструмента для их извлечения из текстов. Раздел 4 описывает результаты факторного анализа и интерпретацию измерений языковой вариативности. Распределение текстов и FTD в пространстве измерений языковой вариативности представлено в разделе 5. Раздел 6 содержит результаты исследования и обсуждение их применения для решения задачи автоматической жанровой классификации.

2. Многомерный анализ Д. Байбера

Многомерный анализ был впервые использован Дугласом Байбером в (Biber, 1986), а затем подробно описан в работе (Biber, 1988). Данный подход был разработан с целью идентификации наиболее статистически значимых сочетаний языковых признаков в устных и письменных жанрах (регистрах) текстов на английском языке. Анализ включает восемь основных методологических этапов, подробное описание которых представлено в работах (Biber, 1988; Biber et al., 2007), и которые мы реализовали в данном исследовании.

1. Подготовка корпуса.

В случае анализа устного дискурса тексты транскрибируются. Байбер подчеркивает (Biber et al., 2007, с. 8, 18, 263), что корпус, созданный специально для поставленной задачи, должен отражать все жанровое многообразие исследуемой области дискурса.

2. Определение набора языковых признаков для извлечения из текстов.

По мнению Байбера (Biber et al., 2007, с. 263), набор должен быть максимально широким и включать все признаки (лексические классы, грамматические категории и синтаксические конструкции), которым можно дать функциональную интерпретацию.

3. Создание специальной компьютерной программы (тэггер) для автоматической разметки всех признаков в корпусе.
4. Непосредственно разметка корпуса по всем языковым признакам.
5. Отдельная компьютерная программа выполняет подсчет частот размеченных признаков в каждом тексте корпуса.
6. Идентификация статистически значимых сочетаний языковых признаков с помощью факторного анализа.

В результате применения факторного анализа большое количество исходных, наблюдаемых переменных (языковых признаков) сводится к сокращенному набору скрытых переменных, называемых факторами. Обосновывая данный статистический метод, Байбер подчеркивает (Biber

et al., 2007, с. 264), что факторный анализ позволяет исследовать совместную встречаемость языковых признаков в текстах. Если какие-то признаки встречаются часто в одних текстах и редко в других, то они имеют большое значение общей дисперсии, т.к. она измеряет вариацию признака под влиянием всех обусловливающих эту вариацию факторов. Факторный анализ направлен на извлечение тех факторов, которые объясняют наибольший объем общей дисперсии признаков и которые упорядочиваются по убыванию их долей дисперсии (Ким и др., 1989, с. 36).

7. Функциональная интерпретация результатов факторного анализа.

Каждый фактор (или измерение вариации языковых признаков) интерпретируется на основе коммуникативных функций определяющих его признаков (в том числе и с положительными, и с отрицательными нагрузками в случае биполярных факторов).

8. Вычисление факторной оценки (оценки измерения вариативности) каждого текста в корпусе, а также средних значений факторных оценок каждого исследуемого жанра с целью изучения их языковой специфики.

Вычисление факторных оценок всех текстов необходимо для сравнения жанров друг с другом, а также для анализа их распределения в пространстве вариаций языковых признаков.

3. Подготовка и анализ данных

3.1. Описание корпуса

Построение тестового корпуса текстов, на котором проводилось данное исследование, можно разделить на несколько этапов. Сначала было выбрано 266 текстов из Open Corpora (Bocharov et al., 2011). Жанровое многообразие первого корпуса было ограничено, поэтому новая группа аннотаторов должна была добавить максимально разнообразные тексты и разметить их. В качестве источников текстов использовались новостные порталы (например, chaskor.ru, ru.wikinews.org), Википедия и другие онлайн энциклопедии, онлайн журналы (например, popmech.ru, afisha.ru), онлайн библиотеки (lib.ru, wikisource.org), блоги (vk.com, lifejournal.com), форумы (например, forum.hackersoft.ru, litforum.ru), порталы с различными научными и научно-популярными публикациями (например, sci-article.ru, research-journal.org), рекламные сайты (например, avito.ru, sportmaster.ru), ресурсы юридических документов (например, base.garant.ru, consultant.ru), онлайн платформы политической и социальной журналистики (politonline.ru, politikus.ru, nstarikov.ru) и другие. На данном этапе объем корпуса составлял 514 текстов. Корпус был размечен одиннадцатью аннотаторами, по три человека на каждый текст. Процедура сбора корпуса описана в статьях (Sorokin et al., 2014; Lagutin et al., 2015).

Таблица 1: Характеристика используемого корпуса.

Количество текстов: 774
Количество слов: 815118
Количество предложений: 62582
Длина текста в словах: 94 (min), 11365 (max), 437 (median)
Количество текстов < 200 слов: 211
Количество текстов > 200 слов: 563

На последнем этапе корпус был расширен до 774 текстов (были добавлены новости, технические статьи, юридические документы, пропагандистские статьи, художественная литература). Количественные характеристики корпуса представлены в таблице 1. В ходе построения корпуса мы пытались отбирать тексты небольшого объема, максимально однородные с жанровой точки зрения. Это было сделано для упрощения процедуры жанровой классификации с использованием функциональных размерностей (FTD) на начальном этапе эксперимента.

3.2. Жанровая аннотация

Корпус был размечен по 17 жанровым функциональным размерностям. Для определения степени согласия между разметчиками использовался такой критерий, как α Криппендорффа, значения которого превышали 75% на первом этапе эксперимента (Sorokin et al., 2014) и достигали значений 90-95% на последнем этапе, описанном в статье (Lagutin et al., 2015).

С целью установления соответствия между метками функциональных размерностей (их наиболее устойчивыми комбинациями) и жанрами размеченные тексты были кластеризованы. В качестве признаков объектов кластеризации (текстов) выступали значения 17 функциональных размерностей. Детальное описание кластеризации представлено в статье (Lagutin et al., 2015). Некоторые кластеры функциональных размерностей были объединены в классы, так как каждый кластер представлен малым количеством текстов, что делает автоматическую классификацию по таким кластерам невозможной. Разбиение на классы C1-C9 (см. таблицу 2) соотносится с результатами предыдущего исследования, описанного в статье (Sorokin et al., 2014), где все результаты были получены полностью автоматически.

Таблица 2: Классы функциональных размерностей.

Класс	Главные функциональные размерности	Количество текстов	Основная интерпретация
C1	A1, A13	28	Аргументативные тексты
C2	A11	177	Личные блоги
C3	A8	82	Новостные тексты
C4	A9	77	Юридические тексты
C5	A12	77	Рекламные тексты
C6	A14	58	Научные тексты
C7	A16 (-A14)	198	Энциклопедические тексты
C8	A7	52	Инструкции
C9	A4, A16	53	Художественная литература

3.3. Разметка языковых признаков

Первый набор признаков для Многомерного анализа, предложенный Байбером в его работе (Biber, 1988), был составлен и функционально проинтерпретирован на основе опыта других исследователей, занимавшихся сравнением устного и письменного типов дискурса (Chafe, 1982; Chafe and Danielewicz, 1985; O'Donnell, 1974; Blankenship, 1974; Quirk et al., 1972). Для целей данного исследования мы в некоторой степени основывались на лексических, грамматических и лексических признаках, представленных в Приложении 2 к работе (Biber et al., 2007).

Первоначально выбранный нами набор включал 40 языковых признаков, затем он был расширен до 63. При его составлении мы руководствовались такими принципами, как доступность и лингвоспецифичность. Во-первых, признаки должны быть доступны для извлечения при помощи таких инструментов, как морфологическая разметка, небольшие словари и поверхностный парсинг, который представляет собой поиск языкового факта в ограниченных контекстных условиях. Во-вторых, набор признаков, используемых Байбером, составлен для английского языка. Наша задача заключается в том, чтобы поиск отобранных признаков полностью соответствовал особенностям русского языка.

Мы исключили как нерелевантные для русского языка следующие признаки: «Contractions» (например, n't и др.), «Existential there», «Hedges» (речь идет о средствах более уклончивого выражения мыслей, например, at about, something like, more or less и другие), «Stranded preposition» (конструкции типа «the thing I was thinking of», в которых предлог оказался вне свойственной ему составляющей) «Split infinitives» (конструкции типа «He wants to convincingly prove that», где маркер инфинитива to и глагольная форма разделены наречием или группой наречий), а также признак «Pro-verb do» (по мнению некоторых ученых, позицию которых мы разделяем, проформы глагольных групп в русском языке не существуют).

Лингвоспецифичность, то есть соответствие особенностям русского языка, проявляется в способе кодирования признаков, например, возвратных местоимений, или неопределенных местоимений, или модальных значений. Мы отказываемся от некоторых признаков, которые не кодируются в русском языке или их функции распределены по разным языковым средствам, как, например, это происходит с признаком «Hedges», но и в данном случае речь идет о кодировании

уже заданного списка признаков, а не добавлении новых, свойственных только русскому языку. Поиск таких новых, неуниверсальных признаков, являющихся уникальными для конкретного языка, может стать предметом дальнейших исследований.

Нами была разработана программа MDRus Analyser[1] на языке Python для идентификации и подсчета частот 63 языковых признаков в текстах корпуса. Программа принимает на вход морфологически размеченный корпус в xml-формате. Для разметки был использован RFTagger (Schmid and Laws, 2008), точность которого близка к точности инструментов, описанных в (Sharoff and Nivre, 2011) и принимает значение около 95-97%[2]. Словари были составлены с помощью Национального корпуса русского языка (НКРЯ)[3].

Все результаты работы программы MDRus Analyser не подвергались никакой пост-обработке (как предлагается в Biber, 2007, с. 263), так как на следующем этапе исследования планируется применять данную программу к большому объему данных, где проверка результатов ее работы вручную нереализуема. Полученные суммы для каждого признака (кроме таких признаков, как средняя длина слова, средняя длина предложения и TTR) в каждом тексте были поделены на длину соответствующих текстов в словах. В результате мы получили матрицу частот языковых признаков размерностью 774 × 63.

4. Анализ измерений языковой вариативности

4.1. Факторный анализ

Задача факторного анализа в том, чтобы исследовать ненаблюдаемую структуру данных и объяснить корреляции внутри набора наблюдаемых переменных с помощью набора фундаментальных ненаблюдаемых переменных, лежащих в основе этих данных. Такие ненаблюдаемые переменные называются факторы, наблюдаемыми переменными являются языковые признаки. В данном исследовании факторный анализ был реализован с помощью пакета psych языка R.

Для того, чтобы решить, сколько факторов извлечь, мы использовали в первую очередь диаграмму осыпи (Cattel, 1965) и критерий Кайзера. Суть критерия состоит в том, чтобы отбрасывать все факторы, собственное значение которых меньше 1,0 (см. таблицу 3). Эти критерии указывают на наличие семи факторов.

Таблица 3: Результат факторного анализа

	PA3	PA1	PA2	PA4	PA5	PA6	PA7
Собственные значения	4,52	4,21	3,25	2,46	2,16	2,01	1,90
Объясненная доля дисперсии	0,22	0,21	0,16	0,12	0,11	0,10	0,09
Накопленная доля дисперсии	0,22	0,43	0,58	0,70	0,81	0,91	1,00

Для выделения семи общих факторов мы использовали метод повторных главных осей. Для упрощения интерпретации факторов было использовано наклонное, или косоугольное, вращение промакс (мы выбрали данный тип вращения, так как предполагаем возможность корреляции между факторами). Межфакторная корреляция находится в пределах от -0,43 до 0,36. Далее мы получили матрицу факторных нагрузок, которую мы можем проинтерпретировать. Для интерпретации каждого фактора учитываются только языковые признаки с коэффициентами факторных нагрузок меньше -0,3 или больше 0.3. Другие признаки отбрасываются как несущественные.

4.2. Интерпретация измерений языковой вариативности

Согласно принятой гипотезе, все языковые признаки, формирующие один фактор (или измерение языковой вариативности), функционально взаимосвязаны. Интерпретируя их

[1] https://github.com/Askinkaty/MDRus_analyser
[2] http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/
[3] http://www.ruscorpora.ru

коммуникативные функции в составе одного фактора, можно дать более общую интерпретацию и самому фактору.

Языковое измерение D1 (Dimension 1) объединяет вокруг себя почти половину всех языковых признаков. Ментальные глаголы, усилительно-ограничительные частицы, изъяснительные и определительные придаточные могут выражать развернутое мнение говорящего. Указательные местоимения, выступающие в функции проформ именных групп или групп прилагательного, а также местоимения третьего лица можно интерпретировать как средства выражения предметного дейксиса. Пространственный и временной дейксис может выражаться с помощью наречий. Отношение и оценка выражаются с помощью соответствующих вводных слов и конструкций. Местоимения третьего лица и глаголы речи служат для передачи чужой речи. Высокая частотность одушевленных существительных может быть связана с тем, что предметом речи являются другие люди.

Отрицательный полюс измерения D1 включает признаки, обозначающие информационную насыщенность текста, например, существительные и их модификаторы (прилагательные в определительной функции), а также числительные.

Измерение D1 мы будем рассматривать как **аргументативное vs. информативное**. Положительный полюс D1 очень близок к аргументативному полюсу измерения Dimension 2 в работе (Berber Sardinha et al., 2014). Отрицательный полюс D1 имеет сходства с отрицательным полюсом информативного измерения Dimension 1 в работе (Grieve et al., 2010).

Измерение D2 в первую очередь может быть проинтерпретировано как диалогичное (или интерактивное), так как самые большие факторные нагрузки имеют признаки, характерные для диалогичной речи (местоимения первого и второго лица, глаголы речи, частицы, восклицательные предложения). Ряд признаков служит для выражения персонального (личные местоимения), пространственного (наречия места) и временного дейксиса (наречия времени). Директивность выражается с помощью повелительного наклонения, инфинитивов и восклицательных предложений. Ментальные глаголы и оценочная лексика могут служить для выражения отношения к предмету речи.

Например, фрагмент диалога из рассказа А. П. Чехова «Лишние люди», включающий такие признаки измерения D2, как формы местоимений первого лица (*я, мне, меня*), формы местоимений второго лица (*ты, тебе, тебя*), частицы (*да, ну, так*), глаголы речи (*визжать, говорит, врешь, браниться*), ментальный глагол (*думает*), восклицательные предложения («*Ты всегда врешь*!» и др.), оценочная лексика (*дурак, умница, славный*), наречия (*убедительно, горько*), инфинитив (*высечь*).

> "*— Ты всегда врешь! Высечь тебя нужно, свиненка этакого! Я тебе уши оборву! — Да ты* что *бранишься*? *— визжит* Петя. — Что *ты* ко *мне* пристал, *дурак*? *Я* никого *не* трогаю, *не* шалю, слушаюсь, а *ты... сердишься! Ну*, за что *ты меня бранишь*? Мальчик *говорит убедительно* и *так горько* плачет, что Зайкину становится *совестно*. «И правда, за что *я* к нему придираюсь?» *— думает* он. *— Ну*, будет… будет, *— говорит* он, *трогая* мальчика за плечо. — Виноват, Петюха… прости. *Ты* у *меня умница, славный, я тебя* люблю".

Отрицательный полюс измерения D2 очень близок к отрицательному информативному полюсу языкового измерения D1. Самые высокие нагрузки имеют такие признаки, как длинные слова, отглагольные и другие существительные и адъективные прилагательные, то есть все те признаки, которые мы проинтерпретировали как обозначающие информативную насыщенность текста. Также D2 включает все пассивные конструкции, и, хотя их факторные нагрузки значительно ниже, чем у вышеперечисленных признаков, они, наряду с отглагольными существительными, выступают как показатели книжного стиля в рамках данного измерения.

Представленный ниже фрагмент Таможенного кодекса РФ включает большое количество существительных (*товаров, населения, здоровья* и др.), отглагольных существительных (*ввоз, вывоз, соображение, защита* и др.) и адъективных определений (*отдельных, транспортных, государственной* и др.).

(1) "*Ввоз* в *Российскую Федерацию* и *вывоз* из *Российской Федерации отдельных товаров* и *транспортных средств* могут быть *запрещены* исходя из *соображений государственной безопасности, защиты общественного порядка, нравственности населения, жизни* и *здоровья человека, защиты животных* и *растений, охраны окружающей природной среды, защиты художественного, исторического* и *археологического достояния народов Российской Федерации* и *зарубежных стран, защиты права собственности*…"

Языковое измерение D2 мы будем интерпретировать как **интерактивное (или диалогичное) vs. информативное**. D2 в определенной степени близко к «Involved vs. Informational dimension» в работе (Berber Sardinha et al., 2014) и к измерению «Informational vs. Involved» в статье (Biber, 1993a).

Положительный полюс измерения D3 объединяет признаки, которые позволяют рассматривать его как нарративное (особенно, прошедшее время, совершенный вид, наречия времени и каузативные глаголы). На хронологические и причинно-следственные связи событий могут указывать наречия времени и каузальные глаголы соответственно. Деепричастные обороты могу описывать побочные линии событий (Плунгян, 2008, с. 135). Мы будем интерпретировать D3 как **нарративное vs. ненарративное** измерение. Сходное измерение вариации языковых признаков представлено в (Biber, 2004) и обозначено как 'Narrative-focused discourse'.

Измерение D4 включает признаки, которые мы уже наблюдали в составе D1 и D2, а также отношение числа лемм к числу словоупотреблений (TTR), обозначающее словарное богатство текста. Пассивные конструкции, длинные слова и отглагольные существительные, имеющие самые высокие факторные нагрузки в D4, подчеркивают его книжный характер. Мы будем трактовать все эти признаки как обозначающие абстрактность текста, а измерение D4 как **абстрактное vs. неабстрактное**.

Измерение D5 также имеет только положительный полюс. Такие признаки, как оценочная лексика, наречия и прилагательные, характеризующие предмет с точки зрения его физических свойств, а также характеристика размера и наречия степени позволяют рассматривать данное измерение как **оценочное vs. неоценочное**.

Измерение D6 включает три признака с отрицательными факторными нагрузками. Все эти признаки объединяет общее понятие количества, в том числе и существительные, обозначающие счет времени (например, день, тысячелетие, эра). Данное измерение мы будем рассматривать как **неколичественное vs. количественное**.

Большинство признаков измерения D7 выполняют директивную функцию, то есть побудить читателя к действию (повелительное наклонение, модальность необходимости) и урегулировать его поведение (модальность необходимости и возможности, придаточные условия, инфинитив, отрицание). D7 мы будем рассматривать как **директивное vs. недирективное** измерение. В некоторой степени оно схоже с измерением 'Overt expression of persuasion' в работе (Biber, 1993a).

Итоговый набор полученных языковых измерений представлен в таблице 4.

Таблица 4: Интерпретация результатов факторного анализа

D1: аргументативность vs. информативность
Положительные признаки: все наречия, указательные местоимения в функции проформ именных групп, частицы, глаголы существования, ментальные глаголы, наречия времени, отрицание, неопределенные местоимения, изъяснительные придаточные, глаголы речи, прошедшее время, местоимения третьего лица, возвратные местоимения, прилагательные в предикативной функции, одушевленные существительные, местоимения *это/этот/тот* как определители именных групп, условное наклонение, вводные слова и конструкции со значение оценки и отношения, относительные придаточные.
Отрицательные признаки: адъективное определение, числительные, существительные, отглагольные существительные, страдательные причастные обороты, средняя длина слова.
D2: интерактивность (диалогичность) vs. информативность
Положительные признаки: местоимения первого лица, отрицание, местоимения второго лица, все наречия, частицы, ментальные глаголы, наречия места, неопределенные местоимения, повелительное наклонение, инфинитив, наречия времени, глаголы речи, оценочная лексика, сочинение клауз, каузативные глаголы, глаголы движения, восклицательные предложения.
Отрицательные признаки: средняя длина слова, существительные, адъективные определения, отглагольные существительные, пассивные конструкции без агента, страдательные причастные обороты, действительные причастные обороты, пассивные конструкции с выраженным агентом
D3: нарративность vs. ненарративность
Положительные признаки: совершенный вид, прошедшее время, каузативные глаголы, глаголы движения, глаголы речи, местоимения третьего лица, наречия времени,

деепричастия, глаголы существования.

D4 : абстрактность vs. неабстрактность

Положительные признаки: страдательные причастные обороты, пассивные конструкции без агента, средняя длина слова, отглагольные существительные, пассивные конструкции с агентом, адъективные определения, действительные причастные обороты, существительные, TTR.

D5: оценочность vs. неоценочность

Положительные признаки: оценочная лексика, все наречия, прилагательные, характеризующие физические свойства объектов, наречия степени, прилагательные со значением размера, длина предложений.

D6: неколичественность vs. количественность

Отрицательные признаки: существительные с количественным значением, числительные, существительные со значением времени

D7: директивность vs. недирективность

Положительные признаки: инфинитив, придаточные условия, модальные слова со значением необходимости, модальные слова со значением возможности, повелительное наклонение, местоимения второго лица, оценочная лексика, отрицание.

5. Распределение классов FTD в пространстве языковых измерений

Следующим этапом является изучение того, как классы функциональных размерностей (FTD) представлены в пространстве измерений языковой вариативности. Мы вычислили значения всех факторных оценок для каждого текста путем сложения частот положительных признаков с наиболее значимыми факторными нагрузками и вычитания частот отрицательных признаков с наиболее значимыми факторными нагрузками. Например, чтобы вычислить значение факторной оценки для некоторого текста по измерению D2, мы суммируем частоты признаков с нагрузками > 0,3 для данного фактора и вычитаем частоты признаков с нагрузками < -0,3 для данного фактора.

Таблица 5: Медианы факторных оценок для C1-C9

Класс	D1	D2	D3	D4	D5	D6	D7	Основная интерпретация класса
C1	7,25	2,62	0,45	-6,06	-1,20	1,79	0,14	Аргументативные тексты
C2	13,69	14,57	5,74	-15,71	3,36	3,49	2,77	Личные блоги
C3	-1,31	-6,11	1,47	3,34	-3,92	-4,80	-6,19	Новостные тексты
C4	-19,03	-21,73	-13,23	25,09	-8,15	-3,71	-3,65	Юридические тексты
C5	-12,52	-7,20	-6,58	9,18	2,13	0,82	1,67	Рекламные тексты
C6	-10,01	-12,22	-6,94	14,34	-2,54	-1,13	-3,09	Научные тексты
C7	-8,76	-9,76	-4,99	9,95	-1,10	-1,99	-3,57	Энциклопедические тексты
C8	-6,08	6,07	1,39	-1,68	2,11	0,01	8,23	Инструкции
C9	13,55	18,50	11,20	-17,51	2,17	4,54	2,00	Художественная литература

Далее мы должны вычислить значения факторных оценок для классов функциональных размерностей C1-C9. Каждый класс имеет одну (или более) главную размерность FTD, определяющую его интерпретацию (см. таблицу 2). Для вычисления факторных оценок каждого класса мы берем факторные оценки текстов, которым было приписано значение 2 по значимым для этих классов функциональным размерностям. Например, для вычисления факторных оценок для класса C4 мы учитываем только оценки текстов, которым было присвоено значение 2 по A9 («До какой степени автор уделяет внимание эстетике текста?»). Медианы факторных оценок для всех классов представлены в таблице 5.

Класс C1 преимущественно включает тексты на социальную, политическую и религиозную тематику. Большинство текстов данного класса аргументативны, диалогичны, нейтральны по измерению «нарративность» и имеют низкие оценки по измерению «абстрактность», а также для них не характерна высокая информационная насыщенность и оценочность. Также большинство текстов не директивны (за исключением религиозных).

Таблица 6: Стандартные отклонения значений факторных оценок классов C1-C9

Класс	D1	D2	D3	D4	D5	D6	D7	Основная интерпретация класса
C1	10,72	10,29	7,40	11,06	2,54	3,08	3,52	Аргументативные тексты
C2	9,93	11,30	6,18	10,46	3,91	4,43	5,84	Личные блоги
C3	7,80	7,41	5,13	7,85	4,11	4,07	4,49	Новостные тексты
C4	7,49	8,08	5,12	8,92	2,98	3,52	4,19	Юридические тексты
C5	11,71	11,81	5,59	13,65	4,51	5,14	5,71	Рекламные тексты
C6	9,06	5,56	3,78	6,89	2,93	3,31	2,46	Научные тексты
C7	12,28	11,30	8,12	13,17	5,64	5,17	5,10	Энциклопедические тексты
C8	10,90	11,92	5,18	12,02	5,46	4,50	7,39	Инструкции
C9	8,72	11,09	7,61	9,06	3,48	5,31	7,63	Художественная литература

Высокие значения среднеквадратичного отклонения (см. таблицу 6) для измерений D1-D4 показывают большой разброс значений факторных оценок и значительное отклонение от средних значений. Анализ класса C1 показал, что отклонения вызваны рядом религиозных текстов, содержащих развернутое обоснование позиций автора по различным религиозным вопросам и текстами о политической ситуации в России.

Класс C2 объединяет личные блоги, которые можно охарактеризовать как аргументативные, диалогичные, директивные тексты, выражающие авторскую оценку. Значительное количество рецензий и отзывов в блогах имеют самые высокие факторные оценки по измерению D1 наряду с высокими оценками по оценочному измерению D5. Мы можем использовать этот факт для идентификации отзывов среди других блогов. Высокие значения среднеквадратичного отклонения по D1-D4 для класса C2 вызваны рядом текстов, которые нельзя определить как личные блоги. Например, это разрешения на обработку персональных данных или медицинское вмешательство.

C3 (класс новостей) включает информативные, недиалогичные и недирективные тексты, для которых, как правило, не свойственна оценка освещаемых событий, а также тексты данного класса имеют относительно высокие оценки по измерениям «абстрактность» и «нарративность». Класс состоит из двух типов новостных текстов: во-первых, это короткие и информационно насыщенные отчеты о событиях с отрицательными оценками по D1, во-вторых, это тексты с развернутой аргументацией позиции журналиста по какому-то вопросу. Некоторые тексты, отмеченные как новостные, содержательно ничем не отличаются от личных блогов. Как следствие, гетерогенность класса C2 отражается на значениях факторных оценок и высоких значениях среднеквадратичного отклонения для измерений D1-D2 и D4 (см. таблицу 6).

В класс C4 (юридические тексты) входят неаргументативные, недирективные тексты, с высокими факторными оценками по измерениям «информативность», «абстрактность» и «количественность», и низкими — по измерению «оценочность». Сравнив факторные оценки по D1 («аргументативность vs. информативность») и D2 («интерактивность vs. информативность»), мы можем сделать вывод, что C4 является самым информационно насыщенным среди всех жанровых классов в нашем наборе.

Большинство текстов класса C5 (реклама) информативны, директивны и выражают оценку описываемых объектов и услуг. В составе текстов, факторные оценки которых в наибольшей степени отклоняются от медианных значений, выделяются две различные группы. Первая группа — объявления с сайтов знакомств. Авторы стремятся максимально привлечь внимание потенциальных партнеров. Вторая группа текстов представляет собой описания различных технически сложных товаров (камеры, автомобили, синтезаторы и т.д.), эти тексты очень информативны и не направлены на взаимодействие с читателем. Высокие оценки класса C5 по измерению D4 («абстрактность vs. неабстрактность») в первую очередь обусловлены второй группой рекламных текстов.

Класс C6 включает научные тексты, которые преимущественно информативны (название понятий преобладает над названием действий, что объясняет высокую частотность существительных), неаргументативны, ненарративны, недирективны, а также для текстов этого класса характерна объективность, абстрактность и высокая частотность чисел и слов с

количественным значением. Мы полагаем, что большое значение среднеквадратичного отклонения по измерению D1 вызвано рядом текстов по филологической тематике, характерные для них признаки смещают оценки в аргументативный полюс.

Класс C7 объединяет в первую очередь энциклопедические статьи. Большинство текстов данного класса информационно очень насыщенные, неаргументативные, абстрактные, ненарративные, не выражающие оценки и содержащие много слов с количественным значением и чисел.

Класс C8 (инструктивные тексты) имеет самые высокие оценки по измерению D7 («директивность vs. недирективность»). В данном подмножестве выделяется две группы текстов: во-первых, это технические инструкции, руководства пользователей, рецепты и другие тексты, которые можно охарактеризовать как информативные и абстрактные; во-вторых, это различные советы, например, как бросить курить. Вторая группа включает тексты, направленные на взаимодействие с читателем и содержащие аргументированное обоснование авторского мнения.

Интересно, что инструкциям по охране труда и медицинским инструкциям характерны низкие факторные оценки по директивному измерению, так как в них отсутствуют типичные для других текстов этого класса формы повелительного наклонения, инфинитивы и средства выражения модальности возможности или необходимости. В нашем корпусе в данных текстах скорее преобладают существительные и формы настоящего времени безличных глаголов. В статье (Berber Sardinha et al., 2014) также подчеркивается, что инструкции к медицинским препаратам не маркированы на инструктивном измерении.

В класс C9 входит художественная литература (преимущественно, это стихи и короткие рассказы). Тексты этого класса диалогичны, нарративны, директивны и содержат выражение оценки. Директивность некоторых текстов класса связана с характером диалогов в рассказах, а также частотными обращениями к читателям в стихотворениях. Среднеквадратичные отклонения по D1-D4 для класса C9 можно объяснить высокой частотой существительных и адъективных определений, а также относительно большими значениями TTR в стихотворениях и песнях (т.е. количеством уникальных лемм). Примечательно, что по низким факторным оценкам по измерению D1 мы можем найти почти все стихи и песни в нашем подкорпусе художественной литературы.

6. Заключение

Мы применили разработанный Д. Байбером Многомерный анализ к жанрово разнообразному корпусу текстов из русскоязычного Интернета. Нами был выбран набор из 63 языковых признаков, построенный на основе признаков, описанных в книге Байбера 1988 года. Набор был адаптирован, так как мы учитывали специфичность кодирования всех признаков в русском языке. В связи с богатством морфологии в большинстве случаев мы столкнулись с большей сложностью правил для идентификации языковых признаков, чем описывается в книге Байбера (Biber, 1988).

Мы применили разведочный факторный анализ к матрице частот языковых признаков, чтобы исследовать их распределение в текстах разных жанров. Далее мы вычислили факторные оценки классов функциональных размерностей, которые рассматриваем как жанры, и рассмотрели распределение этих классов в пространстве полученных измерений языковой вариативности, сделав ряд практически полезных наблюдений. Например, факторные оценки по измерению «оценочность» позволяют различать сходные по факторным профилям аргументативные тексты и блоги. Мы также рассмотрели, как классы функциональных размерностей C1-C9 варьируют внутри. Анализируя значения среднеквадратичного отклонения факторных оценок всех текстов в пределах класса, мы можем обнаруживать в составе данного класса отдельные жанры (или поджанры). Например, в составе инструктивных текстов выделяются технические инструкции и советы; по низким оценкам в измерении «аргументативность vs. информативность».

В будущем планируется проверить релевантность полученных оценок по измерениям языковой вариативности на неоднородных с жанровой точки зрения текстах (например, текстах блогов с комментариями). На данном этапе мы пытались отбирать максимально гомогенные тексты небольшого объема, поэтому фактор неоднородности при реализации Многомерного анализа не учитывался. Так как мы полагаем, что наборы функциональных размерностей FTD могут быть использованы для автоматической идентификации жанров в большом корпусе, необходимо исследовать связь данных размерностей и соответствующих измерений языковой вариативности на объеме большого корпуса. Мы полагаем сделать это следующим шагом нашего исследования.

Литература

1. Бахтин М. М. Проблема речевых жанров. // Литературно-критические статьи. — М.: Художественная литература, 1986, с. 428–472.
2. Ким, Дж.-О., Мьюллер, Ч. У., Клекка, У. Р., Олдендерфер, М. С., Блэшфилд, Р. К., Факторный, дискриминантный и кластерный анализ. Пер. с англ. / Под ред. И. Е. Енюкова. — М.: Финансы и статистика, 1989.
3. Adamzik K. Textsorten – Texttypologie. Eine kommentierte Bibliographie. – Münster, 1995.
4. Berber Sardinha, T., Kauffmann, C. and Mayer Acunzo, C. A multi-dimensional analysis of register variation in Brazilian Portuguese. // Corpora, vol. 9, no. 2, 2014, pp. 239–271.
5. Biber, D. Spoken and written textual dimensions in English: Resolving the contradictory findings. // Language, vol. 62, 1986, pp. 384–414.
6. Biber, D. Variation across speech and writing. Cambridge: CUP, 1988.
7. Biber, D. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. // Computers and the Humanities, vol. 26, 1993a, pp. 331–345.
8. Biber, D. Conversation text types: a multi-dimensional analysis. / Purnelle, G., Fairon, C. and Dister, A. (eds.) // Le poids des mots: Proc. of the 7th International Conference on the Statistical Analysis of Textual Data, Louvain: Presses universitaires de Louvain, 2004, pp.15–34.
9. Biber D., Connor, U. and Upton, T. Discourse on the move: using corpus analysis to describe discourse structure. Amsterdam – Philadelphia, 2007, pp. 261–271.
10. Blankenship, J. A linguistic analysis of oral and written style. // Quarterly Journal of Speech, no. 48, 1962, pp. 419–422.
11. Bocharov, V., Bichineva, S., Granovsky, D. et al. Quality assurance tools in the OpenCorpora project. // Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo, 2011, pp. 101–110.
12. Chafe, W. Integration and involvement in speaking, writing, and oral literature. / D. Tannen (ed.) // Spoken and Written Language: Exploring Orality and Literacy, Nordwood, 1982, pp. 35–54.
13. Chafe, W. and Danielewicz, J. Properties of spoken and written language. / Rosalind Horowitz and S. J. Samuels (eds.) // Comprehending oral and written language. New York: Academic Press, 1985.
14. Forsyth, R. and Sharoff S. Document dissimilarity within and across languages: a benchmarking study. // Literary and Linguistic Computing, vol. 29, 2014, pp. 6–22.
15. Grieve, J., Biber, D., Friginal, E., and Nekrasova, T. Variations among blogs: A Multi-Dimensional Analysis. / Mehler, A., Sharoff, S. and Santini, M. (eds.) // Genres on the Web: Computational Models and Empirical Studies, Berlin – New York: Springer, 2010, pp. 303–323.
16. Lagutin, M., Katinskaya, A., Selegey, V., Sharoff, S., and Sorokin, A. Automatic classification of Web texts using Functional Text Dimensions. // Proc. Dialogue, Russian International Conference on Computational Linguistics, Moscow, 2015 pp. 398–414.
17. O'Donnell, Roy C. Syntactic differences between speech and writing. // American Speech, no. 49, 1974, pp. 102–110.
18. Piperski, A., Belikov, V., Kopylov, N., Selegey, V., and Sharoff, S. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation' // Proc. 8th Web as Corpus Workshop (WAC-8), 2013, pp. 24–29.
19. Santini, M., Mehler, A., and Sharoff, S. Riding the Rough Waves of Genre on the Web. / Mehler, A., Sharoff, S. and Santini, M. (eds.) // Genres on the Web: Computational Models and Empirical Studies, Berlin – New York: Springer, 2010, pp. 3–32.
20. Sorokin, A., Katinskaya, A., and Sharoff, S. Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus. // Proc. Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2014, pp. 646–659.
21. Schmid, H. and Laws, F. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. // Proc. of the 22nd International Conference on Computational Linguistics (COLING'08), Manchester, 2008, pp. 777–784