С.Н. Андреев

КОРПУСЫ СТИХОТВОРНЫХ ТЕКСТОВ И ИХ ИССЛЕДОВАНИЕ МЕТОДОМ ДИСКРИМИНАНТНОГО АНАЛИЗА

Одной из основных целей корпусной лингвистики является такая разметка речевых текстов, которая позволила бы получить информацию о распределении в них признаков, использовать сведения о вариативности признаковых переменных для классификации текстов и/или их частей.

В этой статье ставятся следующие задачи: рассмотреть целесообразность создания корпусов стихотворных текстов; описать применяемый нами способ разметки таких текстов; продемонстрировать возможности применения дискриминантного анализа на базе аннотированного корпуса стихотворных текстов.

Стихотворный текст, как многократно подчеркивалось исследователями, имеет сложный многоаспектный характер, отличается высокой степенью горизонтальной и вертикальной интегрированности по сравнению с текстом прозаическим[1]. Он характеризуется сложным взаимодействием различных уровней, отражающим как общие закономерности языка, так и чисто специфические планы (ритмику, строфику, фонику). Последние накладывают ограничения на текст, требуя соблюдать определенные метроритмические нормы, особую организованность на фонетическом, синтаксическом, лексическом уровнях[2].

---

[1] *Гаспаров М.Л.* Современный русский стих: Метрика и ритмика. М., 1974.

[2] *Баевский В.С.* Пастернак – лирик: Основы поэтической системы. Смоленск, 1993; *Гаспаров М.Л.* Указ. соч.; *Гаспаров М.Л.* Лингвистика стиха // Славянский стих: Стиховедение, лингвистика и поэтика. М., 1996. С. 5–17.

В отличие от прозаической речи стихотворный текст дробится на сопоставимые между собой единицы (строки, строфы) и обладает внутренней мерой – метром. Такое членение на соизмеримые отрезки является обязательным и единообразно заданным.

Стихотворный текст обладает большой насыщенностью информации, которая может быть учтена на различных уровнях. Здесь особенно актуальной становится задача выявления взаимодействия между языковыми и художественными (стихотворными) формами, установление взаимосвязи различных языковых и стихотворных планов.

Обладая большой насыщенностью информации, стихотворный текст в то же время является достаточно компактным по размерам. Так, лирические стихотворения, как правило, не превышают 80 строк. Учитывая такую компактность лирических текстов, а также то, что в лирике в наибольшей степени проявляется индивидуальный стиль авторов, мы сочли целесообразным ориентироваться в первую очередь на лирические тексты.

Разметка стихотворных текстов делает необходимым привлечение наравне с морфологическими, синтаксическими, фонетическими также и ряда специфических для стихотворного текста характеристик – ритмо-метрических, рифменных, строфических признаков, признаков стихотворного синтаксиса.

В рамках коллективной темы по многомерному анализу языковых и речевых единиц на кафедре иностранных языков Смоленского государственного педагогического университета проводится создание корпуса стихотворных лирических текстов американских и английских поэтов-романтиков.

Необходимость при разметке приписывать большое число признаков для ограниченного по размерам текста делает целесообразным построение соотнесенной с текстом таблицы данных. Каждая строка стихотворного текста заменяется строкой, отражающей параметры текста, описывающие его на различных уровнях. При этом мы ставили задачу, чтобы эти параметры позволяли в дальнейшем учитывать как отдельные, изолированные эле-

менты строки, так и различные комплексы признаков, отражающих горизонтальную интеграцию строки, вертикальную интеграцию текста.

Далее представлены строки из произведения Лонгфелло и фрагмент таблицы (табл. 1), отражающий эти строки.

«My plumage bears the crimson blush,
When ocean by the sun is kissed;…»

(Лонгфелло «The Sea Diver»)

В электронном виде эта таблица строится в рамках программы Excel. Для каждой строки указывается автор, название произведения, номер строфы (если стихотворение не астрофическое) и номер строки. В приводимом фрагменте аннотации эти данные показаны в столбцах 1–4.

Общее количество «базовых» признаков, т.е. признаков, непосредственно отраженных в таблице – свыше шестидесяти, из которых здесь представлены лишь некоторые. Для обработки полученных результатов статистическими методами необходимо учитывать характер используемых признаков, являются ли они качественными или количественными.

Количественные признаки представляют собой измеряемые или исчисляемые количества, для их значений осмысленна операция сравнения (на или во сколько раз одно значение больше другого). Качественные признаки, в отличие от них, не поддаются измерению. Им соответствуют порядковые и номинальные шкалы измерения. Порядковые признаки определяются упорядоченным рядом состояний (например, низкая/средняя/высокая словообразовательная продуктивность). Порядок состояний имеет смысл, но интервалы между ними не заданы. Номинальные признаки характеризуются неупорядоченным рядом состояний. В этом случае признак приписывает объектам лишь имена. Частным случаем таких качественных номинальных признаков являются дихотомические признаки, имеющие только два состояния: наличие/отсутствие.

Таблица 1. Фрагмент разметки строки из стихотворения Лонгфелло «The Sea Diver»)

Признаки																						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Автор	Произведение	№ строфы	№ строки	Часть речи начало стиха	Часть речи конец стиха	Слоги начало стиха	Слоги конец стиха	Перенос-начало	Перенос-конец	Разрыв	Тип рифмы (по точности)	Вид рифмы	Икт 1	Икт 2	Икт 3	Икт-конец	Анакруса	Отягчение анакр.	Горизонт.протяженность	Кол-во строф в тексте	Кол-во типов строф	Тип строфы
L	L-12	2	5	n	n	2	1	–	–	–	e	m	–	–	–	–	1	–	4	8	1	4
L	L-12	2	6	n	v	2	1	–	–	–	e	m	–	+	–	–	1	–	4	8	1	4

Среди используемых при разметке признаков количественные составляют лишь небольшую часть. К ним относятся, например, такие как количество стоп в строке (размер), количество предложений (простых, сложносочиненных, сложноподчиненных) в строфе (в строке), количество слогов в слове, занимающем ту или иную позицию в строке и др. Кроме того, количественными являются некоторые параметры всего стихотворного текста: количество строк, количество строф в тексте, количество видов строф и др.

Однако большая часть характеристик, принятых в лингвистике и стиховедении, являются качественными признаками. К ним относятся, например, морфологическая характеристика слова, тип его синтаксической функции, семантические признаки, большинство стихотворных признаков (наличие стихотворного переноса, разрыва строки синтаксической паузой, характеристика рифмы, пропуски ударения на сильных местах стиха и многие другие).

Для применения некоторых статистических процедур необходимо представлять признаки в одних случаях в виде качественных дихотомических, в других случаях – в виде количественных характеристик. В связи с этим полученная в результате аннотации текста таблица, содержащая, как указывалось выше, комбинацию количественных, качественных дихотомических признаков и качественных признаков с более чем двумя состояниями, должна давать возможность перехода к таблицам «объект-признак» с (А) дихотомическими и (Б) количественными признаками.

(А) При преобразовании исходной таблицы в таблицу с дихотомическими признаками необходимо вносить вместо каждого количественного или качественного признака с рядом неупорядоченных состояний несколько качественных дихотомических характеристик, отражающих интервал (для количественных признаков) либо каждое состояние (для качественных признаков). Так, признак «количество стоп в строке» представляется как ряд

признаков – «наличие в строке одной стопы», «наличие в строке двух стоп», «наличие в строке трех стоп» и т.д. Признак «частеречная принадлежность слова в последней сильной позиции» разбивается на ряд дихотомических признаков «наличие существительного в последней сильной позиции», «наличие глагола в последней сильной позиции», «наличие прилагательного в последней сильной позиции» и т.д. Эти преобразования довольно просто алгоритмизируются в рамках программы Excel.

(Б) Для преобразования качественных признаков в количественные мы учитываем частотность значения признака (например, частотность появления существительных в последней сильной позиции) в рамках определенной протяженности текста (строфы, произведения).

Еще одной важной задачей, которая ставится в процессе работы с исходной БД, является вывод новых признаков путем обобщения и объединения признаков базовых. Исходная БД позволяет генерировать большое количество новых признаков.

Так, когда на определенном этапе исследования стало выясняться, что часть речи в последней сильной позиции строки обладает достаточно большой дискриминантной силой, стало желательным детализировать этот признак, комбинируя его с типом рифмы (точная/неточная, мужская/женская и т.д.), с наличием/отсутствием переноса, с длиной строки и т.д. Были выведены новые «комплексные» признаки.

Создание указанного выше корпуса стихотворных текстов позволяет решить ряд классификационных задач с применением дискриминантного анализа.

Дискриминантнгый анализ – один из наиболее подходящих методов для решения классификационных задач при работе с априорно заданными классами. Это статистический метод, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно и решает во-

просы интерпретации межгрупповых различий, а также классификации новых наблюдений по группам[1].

Одним из участников коллективной темы при помощи дискриминантного анализа была проведена классификация лирических произведений американских авторов – Брайента, Лонгфелло, Эмерсона и По[2].

Степень сходства между классами текстов указанных авторов приводится в табл. 2.

Таблица 2. Квадрат расстояния Махаланобиса между центроидами классов текстов Брайента, Лонгфелло, Эмерсона и По

Классы Текстов	**Класс 1 (Брайент)**	**Класс 2 (Лонгфелло)**	**Класс 3 (Эмерсон)**	**Класс 4 (По)**
Класс 1 (Брайент)	0,00	9,11	23,39	29,06
Класс 2 (Лонгфелло)	9,11	0,00	16,15	19,61
Класс 3 (Эмерсон)	23,39	16,15	0,00	22,90
Класс 4 (По)	29,06	19,61	22,90	0,00

В табл.2 отражено расстояние между центрами классов текстов в пространстве признаков. Чем больше расстояние между центрами классов, тем больше различаются эти классы. Под центром класса понимается точка в пространстве с координатами, которые являются средними значениями переменных всех объектов в данном классе. Для определения расстояния используется мера Махаланобиса.

---

[1] *Клекка У.Р.* Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ. М., 1989. С. 78–138.

[2] *Андреев В.С.* Классы стихотворных текстов (на материале лирики американских поэтов-романтиков). Автореф. дисс. … канд. филол. наук. Смоленск, 2002.

Соотношение классов, представленное в табл. 2, можно отобразить на плоскости в виде схемы, которая изображена на рис. 1.

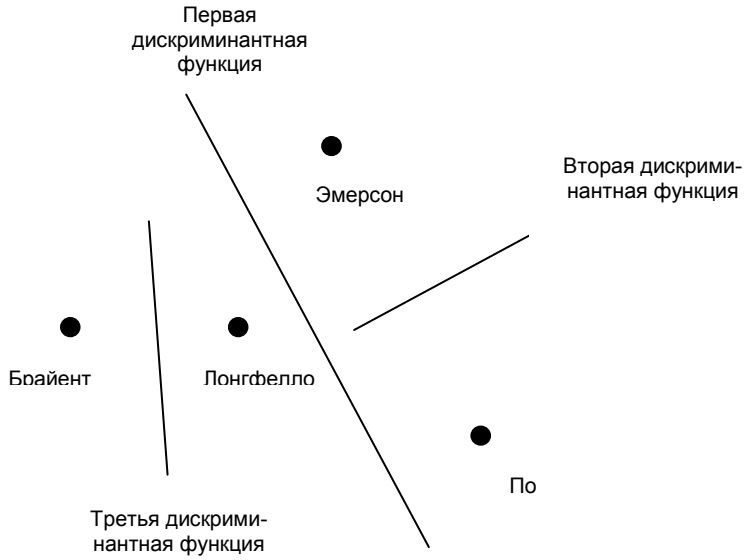

Рис. 1. Схема разграничения классов текстов дискриминантными функциями

Традиционно считалось, что Эмерсон и По являлись самобытными американскими поэтами, а Лонгфелло и Брайент следовали принципам европейской, в первую очередь английской, поэтической культуры, причем в наиболее явном виде имело место противопоставление творческой манеры Лонгфелло, с одной стороны, и Эмерсона и По, – с другой.

Результаты анализа в целом хорошо соотносятся с этим мнением: классы текстов сгруппированы в два противопоставленных кластера (кластер Брайента-Лонгфелло и кластер Эмерсона-По).

В то же время, полученные результаты позволяют уточнить эту классификацию. Так, класс произведений Лонгфелло занимает «центральное» положение относительно всех рассматриваемых авторов. Наблюдается максимальное противопоставление не Лонгфелло, с одной стороны, и Эмерсон и По – с другой, а Брайента относительно Эмерсона и По. Лонгфелло ближе к Брайенту, но также достаточно близок к Эмерсону и По.

В ходе анализа было получено 3 статистически значимые дискриминантные функции. Каждая функция разграничивает определенные классы текстов. Признаки, являющиеся переменными этих функций, имеют коэффициенты, которые позволяют нам судить о вкладе каждого признака в дискриминацию.

На рис. 1 показано, как три дискриминантные функции разграничивают классы текстов. Первая, самая сильная дискриминантная функция, делит классы текстов на два основных кластера. В один входят произведения Лонгфелло и Брайента, в другой – По и Эмерсона. Это, как было сказано, хорошо согласуется с традиционным подходом.

Однако, если традиция основывается на учете тем, сюжетов и, отчасти, образов в произведениях, то мы определили чисто формальные параметры, дифференцирующие эти классы. Ими являются рифменные, строфические и ритмические признаки, такие как количество видов строф, мужских рифм и количество внесхемных ударений на анакрусе.

Вторая и третья функции разделяют классы текстов в рамках двух основных кластеров. В отличие от первой функции здесь более важными являются морфологические и синтаксические признаки. Причем для кластера Эмерсона-По – только морфологические (количество слов различных частей речи в последней сильной позиции), а для кластера Брайента-Лонгфелло – не только морфологические, но и синтаксические, такие как количество разрывов строки синтаксической паузой, количество подлежащих в первой сильной позиции и другие.

Теперь рассмотрим пример, в котором решается задача идентификации объекта в пространстве имеющихся классов.

Творчество Эмили Дикинсон занимает совершенно особое место среди американских поэтов. С одной стороны, Дикинсон считается одним из выдающихся, наиболее ярких американских поэтов-романтиков. С другой стороны, серьезной проблемой является отнесение ее творчества к тому или иному направлению в рамках американского романтизма.

Была поставлена задача выяснить, с произведениями какого из двух выделенных основных кластеров тексты Дикинсон более сходны, если сходны вообще.

В качестве материала для исследования были взяты 20 лирических произведений Дикинсон, отобранных методом случайной выборки и отражающих два этапа ее творчества: ранний и зрелый.

Результаты анализа представлены в табл. 3, в которой дается вероятность отнесения каждого текста к тому или иному классу. В первом и втором столбце даны ***номер*** и название произведения, в остальных столбцах – вероятность отнесения данного текста к соответствующему классу. Как видно из табл. 3, творчество Дикинсон характеризуется следующими основными тенденциями.

В первом периоде (тексты 1–10): к кластеру Брайента-Лонгфелло относится три текста Дикинсон, к кластеру Эмерсона-По – шесть текстов. Текст 1 занимает промежуточное положение: он одновременно сходен с классами Эмерсона и Лонгфелло. Текст 3 занимает аналогичное промежуточное положение внутри кластера Эмерсона-По.

Во втором периоде (тексты 11–20) в кластер Эмерсона-По входит девять текстов (тексты 11–14 и 16–20), в кластер Брайента-Лонгфелло – один текст (текст 15). На основании полученных данных можно сделать следующий вывод. Первый период творчества Дикинсон более «диффузен» с точки зрения модели, образца. Второй период характеризуется значительным следованием манере, отраженной в текстах Эмерсона и По.

Таблица 3. Вероятности отнесения текстов Дикинсон
к классам текстов других исследуемых американских поэтов-романтиков

№	Тексты Дикинсон	Класс Брайента	Класс Лонгфелло	Класс Эмерсона	Класс По
Первый период творчества					
1	Success	0,00	**0,41**	**0,58**	0,00
2	Almost	0,00	**0,90**	0,10	0,00
3	«I asked no other thing»	0,00	0,01	**0,47**	**0,52**
4	Dawn	0,00	0,00	**1,00**	0,00
5	«I had no time to hate, because»	0,00	0,00	0,00	**1,00**
6	«If you were coming in the fall»	0,04	**0,91**	0,05	0,00
7	The Wife	0,02	0,17	**0,57**	0,24
8	«Perhaps you'd like to buy a flower»	0,00	0,00	0,00	**1,00**
9	«As children bid the guest good-night»	0,00	0,00	**1,00**	0,00
10	From the chrysalis	**0,68**	0,05	0,26	0,01

Таблица 3. Окончание.

№	Тексты Дикинсон	Класс Брайента	Класс Лонгфелло	Класс Эмерсона	Класс По
Второй период творчества					
11	«Troubled about many things»	0,00	0,00	**1,00**	0,00
12	Refuge	0,00	0,00	**1,00**	0,00
13	«The bustle in a house»	0,00	0,00	**0,84**	0,16
14	«It was too late for man»	0,00	0,00	**1,00**	0,00
15	«Sleep is supposed to be»	0,37	**0,55**	0,08	0,00
16	«Tell all the truth but tell it slant»	0,00	0,00	**1,00**	0,00
17	«The Bible is an antique volume»	0,00	0,00	**1,00**	0,00
18	«Apparently with no surprise»	0,00	0,00	**1,00**	0,00
19	«My life closed twice before its close»	0,00	0,00	**1,00**	0,00
20	«Elysium is as far as to»	0,00	0,00	0,00	**1,00**