Антонова А.Ю., Клышинский Э.С., Ягунова Е.В. Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости // Труды международной конференции «Корпусная лингвистика–2011». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011

А.Ю. Антонова, Э.С. Клышинский (МИЭМ), Е.В. Ягунова (СПбГУ)

ОПРЕДЕЛЕНИЕ СТИЛЕВЫХ И ЖАНРОВЫХ ХАРАКТЕРИСТИК КОЛЛЕКЦИЙ ТЕКСТОВ НА ОСНОВЕ ЧАСТЕРЕЧНОЙ СОЧЕТАЕМОСТИ[1]

1. Введение

Тексты разных функциональных стилей отличаются по частотности синтаксических конструкций[2]. Тем не менее, до настоящего времени не проводилось анализа особенностей синтаксической структуры для текстов разных функциональных стилей, жанров и/или предметных областей. Одной из причин недостаточной изученности этого вопроса оказывается то, что существующие методы синтаксического анализа не позволяют проводить однозначную синтаксическую разметку текстов. Ниже мы предлагаем метод, также не проводящий полный синтаксический разбор текста. В своей работе мы используем

[1] Работа выполнена при поддержке гранта РФФИ № 10-01-00800.

[2] *Герд А.С.* Специальный текст как предмет прикладной лингвистики / А.С. Герд // Прикладное языкознание : учебник / отв. ред. А.С. Герд. – СПб. : Изд-во С.-Петерб. ун-та, 1996. – С. 68-90.

статистический метод извлечения информации о частеречной сочетаемости слов[3].

На данном этапе единицей анализа является коллекция в целом. Были взяты коллекции текстов трех функциональных стилей: художественных, новостных и научных. Объем каждой из коллекций от 0,7 млн до 6 млрд словоупотреблений (с/у). Объем основных коллекция приведен в табл. 1.

Таблица 1. Объемы коллекций

Источник	**Объем (млн. с/у)**
Художественные тексты	
Библиотека Мошкова	688
WebReadings	3000
Librusec	6000
Новостные источники	
Лента.ру, 2005-2010 гг.	42,5
РИА Новости	186,8
РосБизнесКонсалтинг, 2003-2010 гг.	28,8
Независимая Газета, 1999-2010 гг.	100,6
Российская Газета, 2000-2010 гг.	42
Взгляд	72,4
Мембрана	3
Открытые системы	3,6
Смешанные стили и жанры	
ItHappens	0,7
Популярная механика	2,1
Научные тексты различных стилей и предметных областей	

[3] *Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю.* Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010. С. 181-185

Журнал «Информатика и системы управления», 2001-2010 гг.	1
Журнал «Программные продукты и системы», 1998-2010 гг.	2,5
Коллекция авторефератов различных областей науки	31,4
Конференция КИИ	1,1
Конференция RCDL	1,3

Анализ проводился по следующим сочетаниям: глагол + существительное, глагол + наречие, деепричастие + существительное, деепричастие + наречие, причастие + существительное., причастие + наречие., прилагательное + существительное., существительное + существительное. Для подробного анализа были выбраны три параметра: «существительное + существительное», «глагол + существительное» и «деепричастие + наречие». Целью исследования являлась проверка возможности классификации текстов по стилевым характеристикам при помощи частотных характеристик указанных выше параметров.

2. Предварительные результаты. Обсуждение и интерпретация

Первый параметр – «существительное + существительное в род.п.» (генетивная конструкция, часто соответствующая неоднословному термину). Большое количество этих конструкций традиционно рассматривается как, с одной стороны, морфолого-синтаксическая характеристика научных текстов, а с другой стороны – образчик плохо воспринимаемого стиля и поле работы редактора. По результатам экспериментов художественные коллекции содержали конструкции данного типа от 1 до 1,7% от общего числа выделенных конструкций, новостные тексты – от 18 до 29% и научные тексты – от 35 до

43% конструкций. Все коллекции оказались хорошо отделимы друг от друга по данному параметру.

В качестве второго параметра рассматривалась доля конструкций «глагол + существительное». Второй параметр интересен сам по себе и в сопоставлении с первым. Большое количество конструкций «глагол + существительное» характеризует динамические тексты, т.е. тексты, в которых реализуется большое число ситуаций. Этот параметр взаимодополняет второй, т.к. обилие генетивных конструкций, напротив, отличает статические тексты, т.е. тексты, в которых сообщается о некотором положении дел. В первом случае – рассказ о событиях, во втором – называние события. Именно поэтому для восприятия легче тексты с конструкциями «глагол + существительное».

Эксперименты показали, что в художественных текстах выделяется стабильно около 57-58% от всех конструкций данного типа. Новостная лента показывала 31-40% конструкций «глагол +существительное», научные тексты – от 20 до 28%. Таким образом, разделение текстов по жанрам также оказалось хорошим.

В качестве третьего параметра были взяты конструкции «деепричастие + наречие». Такого типа конструкции встречаются в текстах сравнительно редко. Они, вероятно, характеризуют те тексты, в которых делается особый акцент на действиях (оценивается «качество действия»). Можно предположить, что эти конструкции будут маркировать наиболее яркие тексты, реализуя функции воздействия на адресата.

Для художественных текстов процент подобных конструкций составил около 0,5%. Научные и новостные коллекции по данному параметру разделить не удалось. Для научных коллекций параметр принимал значения от 0,02 до 0,09%, для новостей – от 0,06 до 0,1%. Таким образом, по данному параметру отделимы только часть текстов.

В результате анализа промежуточных результатов классификаций, полученных на основании трех простых параметров, был выбран один комбинированный параметр, максимально отражающий соотношение динамичности / статичности текстов коллекции. Четвертый параметр в числителе имеет частеречные сочетания, характеризующие динамичность текста, а в знаменателе – его статичность: «((гл+сущ) + (гл+нар) + (деепр+сущ) + (деепр+нар)) / ((сущ+сущ) + (прил+сущ))».

Эксперименты показали, что в художественных текстах подобное отношение принимает значения от 2,16 до 2,2; в новостных текстах от 0,67 до 0,83 (за исключением текстов смешанной тематики, для которых было получено значение 1,38); для научных текстов были получены значения от 0,29 до 0,53. Таким образом, и здесь жанровое разделение текстов было произведено успешно.

3. Обсуждение результатов. Выводы

Был проведен сравнительно подробный анализ возможностей использования трех простых (некомбинированных) параметров: определение результирующего разбиения на классы, границ классов, коллекций, которые ведут себя неоднозначно по отношению к используемому параметру, и наборе коллекций, представляющих наиболее однородную выборку с точки зрения стиля и жанра.

Было выявлено, что наиболее неоднозначные результаты показывают тексты со смешанным типом. Например, корпус текстов «It happens», содержащий в себе околокомпьютерные истории, чаще всего оказывался на границе между новостями и беллетристикой.

Выявленная неоднозначность может быть интерпретирована двояко:

- рассматриваемая коллекция содержит тексты, разнородные по стилевым характеристикам,
- тексты этой коллекции реализуют смешение стилей (жанровую неоднородность) внутри самого текста.

Разрешить неоднозначность интерпретации можно лишь в результате включения в наше исследование новой единицы анализа, т.е. рассмотрение распределения значений параметров по текстам в рамках указанных коллекций.

Коллекции, представляющие наиболее однородную стилевую выборку для новостных и научных текстов:

- По первому параметру – для новостных текстов это издания «Мембрана», «Компьюлента», «PCWeek», «Российская газета» и «РБК»; для научных – сборник конференции RCDL, журнал «Программные продукты и вычислительные системы», а также подборки авторефератов.
- По второму параметру – для новостных это «Мембрана», «Компьюлента», «Популярная механика»; для научных – подборки авторефератов.

Согласно рассмотренным интегральным параметрам, «Мембрана» и «Компьюлента» формируют однородную по стилю коллекцию. Является ли это характеристикой коллекций в целом? Априори можно было предположить скорее смешение жанров и стилей в текстах, одновременно характеризующихся новостной и научно-популярной природой в рамках текстов этой коллекции. И этот вопрос надо решать через анализ распределений значений параметров по текстам в рамках указанных коллекций.

Также следует заметить, что результаты могут отображать особенности использованного метода. Так, например, для глагола «думать» в научных текстах было выделено гораздо меньше конструкций. Это связано с тем, что в художественных текстах чаще всего употребляются фразы вида «думать что-то о чем-то», тогда как в научных текстах более распространены конструкции

вида «мы думаем, что», которые не могли быть приняты в рассмотрение.

Несмотря на это, метод показал свою пригодность для классификации коллекций текстов по различным жанрам. В дальнейшем мы планируем изучить статистические характеристики отдельных текстов или их фрагментов, распределение частот встречаемости различных конструкций по элементам текста равного размера.

Anne Antonova, Edward Klyshinsky, Elena Yagunova

TEXT COLLECTIONS GENRE AND STYLISTIC CATEGORIZATION BY POS CO-OCCURENCE

Texts belonging to various functional styles and genres differ in the frequency of some special syntactic constructions. In the article a method is proposed that categorizes texts in different functional styles using statistical information of words' part-of-speech co-occurrence. We analyzed over 20 text collections of the three functional styles: fiction, news and research. The volume of each collection: from 0.7 million to 6 billion tokens. The obtained numeric data allows dividing scientific, news and fictional text collections by simple parameters such as noun + noun, verb + noun, adverbial participle + adverb constructions. The choice of these parameters is the result of the preliminary study.