**Статистические данные в корпусных исследованиях научного стиля.**

С.А.Стринюк

Статистические данные в исследованиях научного стиля играют важную роль. Если ранее в условиях отсутствия компьютерных технологий и возможностей машинной обработки текста исследователю были доступны только ручная обработка текста и ручной подсчет частотности языковых единиц, что делало задачу проведения исследования на репрезентативном корпусе практически невыполнимой, то сейчас доступны разнообразные ПО, позволяющие в сжатые сроки получить данные по значительному количеству текстов и получить более убедительные объективные результаты исследования самых разнообразных языковых явлений.

Наблюдения над частотностью использования различных языковых единиц были проведены задолго до появления популярных ныне инструментов анализа языка (AntConc, Wordsmith и многих других). В.В.Виноградов еще в 1938 г писал:

«По-видимому, в разных стилях книжной и разговорной речи… частота употребления разных типов слов различна. Точные изыскания в этой области помогли бы установить структурно-грамматические, а отчасти и семантические различия между стилями… Анализ всех грамматических категорий должен уяснить их относительный вес в разных стилях литературного языка. Но, к сожалению, пока еще этот вопрос находится в подготовительной стадии обследования материала».

Позднее, делая особый акцент на коренном различии языка и речи Л.Р. Зиндер и Т.В. Строева отмечали, что «…при изучении языковой системы… статистика малоэффективна. Напротив, при изучении речи она приобретает первостепенное значение» (Л.Р. Зиндер и Т.В. Строева, 1968).

* Научный стиль — стиль научных сообщений. Сфера использования этого стиля — наука и научные журналы, адресатами текстовых сообщений могут выступать учёные, будущие специалисты, ученики, просто любой человек, интересующийся той или иной научной областью; авторами же текстов данного стиля являются учёные, специалисты в своей области. Целью стиля можно назвать описание законов, выявление закономерностей, описание открытий, обучение и т. п.
* Основная его функция — сообщение информации, а также доказательство её истинности. Для него характерно наличие малых терминов, общенаучных слов, абстрактной лексики, в нём преобладает имя существительное, немало отвлечённых и вещественных существительных.
* Научный стиль существует преимущественно в письменной монологической речи. Его жанры — научная статья, учебная литература, монография, школьное сочинение и т. д.

Стилевыми чертами этого стиля являются подчёркнутая логичность, доказательность, точность (однозначность)

Специфические стилистические методы:

* семантико-стилистический метод исследования текста
* метод «слово-образ (микрообраз)»
* анализ по стилистическим пометам в словарях и справочной литературе
* наблюдение над нейтральными и стилистически окрашенными средствами в речи
* стилистическая интерпретация текста
* метод стилистического эксперимента.

Семантико-стилистический метод исследования текста является основным специфическим методом. Существует проблема адекватности выражения оттенков смысла со стилистическими значениями. На семантико-стилистических связях языковых единиц базируется взаимосвязь языковых средств и общая стилистическая окраска функциональных стилей речи. Методом сравнения (сопоставления) определяется специфика существующих функциональных стилей

Для определения особенностей функционирования языка в речи и определения языковых особенностей отдельного стиля может быть применен инструментарий математической статистики. С его помощью можно описать закономерности функционирования языка в разных сферах общения, типы текстов, специфику функциональных стилей и воздействующих на них различных экстралингвистических факторов.

**Стилостатистический анализ: качественно-количественный подход**

Сочетание количественного подхода с качественным (лингвостилистической стороной языковых единиц и текста) занимается стилостатистический анализ. Он обращается к семантике исследуемых языковых явлений, его помощью можно выявить взаимосвязь этих явлений с мышлением, целями и задачами общения и других экстралингвистических факторов, воздействующих на речь.

Тем самым стилостатистический метод (= методика) анализа является двуединым качественно-количественным инструментом анализа речевых явлений.

Целью применения метода является выведение стилостатистических (вероятностных) закономерностей.

Языку и речи свойственна вероятностность, которая по-разному проявляется в пределах этой дихотомии. В 1963 г. Н.Д. Андреев и Л.Р. Зиндер определили и ввели в научный обиход понятие речевой вероятности. К примеру, если шесть падежей русского языка являются равноправными в системе языка, то поведение их в речи нарушает это равноправие, частота их употребления оказывается различной.

"система речевой вероятности… есть совокупность относительных количественных характеристик, описывающих численные соотношения между элементами (или группами элементов) в некотором массиве текстов. Можно сказать, что речевая вероятность определяет статистическую структуру текстов, тогда как язык характеризуется их теоретико-множественной структурой и алгоритмами их порождения и распознавания" (*Андреев, Зиндер*, 1963)

Основания применения статистических методов при изучении речи

* объективная присущность речи количественных признаков (повторяемости, частотности единиц)
* стремление получать объективные показатели вместо расплывчатых *часто, редко, обычно*,
* по анализу части текста судить обо всем тексте или по части текстов какой-либо сферы общения – обо всей их совокупности.
* С. методы способствуют и определению закономерностей функционирования языка в речи

Результами применения стилостатистического метода является

* стилевая дифференциация речи, в том числе подтверждению данными статистики вначале интуитивно выделенных основных функциональных стилей,
* объективное решение вопроса о внутристилевой дифференциации этих стилей
* выделению подстилей, жанров и т.д.,
* определение идиостиля,
* также определению пограничных и периферийных зон

Применительно к нашему исследованию стилостатистический метод может быть применен на корпусах текстов, собранных по определенных критериям: 1. принадлежности предметно ориентированному текста к научному стилю, 2. принадлежности текста к научноучебному подстилю (т.е. авторство этого текста принадлежит студенту, а цель его работы не информирование о результатах научного исследования, а научноучебная работа (ВКР, КР, реферат и т.д.)

Сущность стилостатистического метода заключается в том, что из генеральной совокупности (напр. всех текстов науч. стиля) выбирается некоторое число выборок (отрезков текста определенной длины либо изучаемых языковых единиц) и по ним с помощью статистической обработки полученных данных делает заключение о всей генеральной совокупности в отношении исследуемого явления.

Предметом стилистического исследования с применением стилостатистического метода является исследование частотности употребления глаголов настоящего времени в научном стиле речи, в том числе разных семантико-грамматических оттенков глагола, соотношения именных и глагольных форм и т.д.

Подобный анализ позволяет выявить интересные *стилостатистические закономерности функционирования* (см.) языка в речи, а также установить и объективно показать качественный характер науч. стиля речи (*Митрофанова*, 1973)

Существует два возможных способа описания речевых явлений, один из них осуществляет систематическое описание всех языковых явлений, другой ограничивает подход к речевым явлениям рамками определенного функционального стиля.

Научный стиль определяется рядом экстралингвистические факторов:

* цели общения в той или иной сфере деятельности
* назначение соответствующей формы сознания
* особенностей типа мышления и др.

Наша задача – выдвинуть гипотезу о наиболее свойственных и, следовательно, достаточно частотных для исследуемых текстов каких-либо языковых формах (с учетом их семантической стороны), лексико-грамматических значениях и подвергнуть вначале их статистическому анализу, т.е. будут выборочно анализироваться некоторые языковые единицы. Именно те, которые предположительно соответствуют специфике научного стиля, особенностям подстиля и т.д. Стилостатистический метод объективно подтвердит (либо откорректирует или отвергнет) выдвинутую гипотезу и определит путь дальнейшей работы.

Лингвистический анализ текста

* Стилистические функции и особенности лексико-фразеологических средств текста.
* Особенности и стилистические функции словообразовательных средств текста.
* Средства создания образности и экспрессивности текста.
* Индивидуальность слога (приемы, отступления от норм и т.п.)

**Опыт сбора данных по тестовому корпусу:**

В качестве примера применения стилостатистического метода нами была выбрана категория abstract semantic verbs – глаголы широкой абстрактной семантики. Корпус, на котором собраны данные, BAWE (British Academic Writtten Corpus), репрезентативный корпус, содержащий 2761 текстов длиной от 500 до 5000 слов, написанных студентами разных специальностей под научным руководством преподавателей университетов. В корпусе представлены тексты 35 различных специальностей, в том числе гуманитарные и естественные науки) четырех уровней подготовки (бакалавров и магистров).

Для анализа были выбраны глаголы широкой абстрактной семантики, поскольку они являются репрезентантами одного из ключевых атрибутов научного стиля – абстрактности и высокой степени обобщения информации, представленной в научном тексте.

Собранные данные позволяют сделать заключение о характере распределения значений. Очевидно, что в большем количестве текстов соотношение глаголов абстрактной семантики ко всем токенам в тексте укладываются в диапазон от 0,2 до 0,4. Вместе с тем, так же очевидно, что некоторое «распыление» значений происходит в большинстве случаев сторону увеличения, а не уменьшения этих средних значений, что требует дальнейшего более детального рассмотрения отдельных текстов и анализа причин этого явления.

*Диаграмма 1. abstract semantic verbs: sentences with occurrence to all sentences in the document*

Та же процедура, произведенная на корпусе, значительно меньшем по размеру (21 файл), наглядно показывает, насколько более размытыми и непоказательными получаются результаты. Исследование было проведено на текстовом корпусе статей по компьютерным наукам, опубликованным в рецензируемых англоязычных журналах высокого уровня. На диаграмме также видно, что приблизительно в трети текстов показатели находятся в диапазоне от 0,2 до 0,3, а диффузия значений происходит на более высоких показателях, что также требует дальнейшего исследования при анализе отдельных текстов: исследования метаданных текста, и подробном анализе все случаев появления глаголов этого типа в исследуемых текстах.

Диаграмма 2. abstract semantic verbs: sentences with occurrence to all sentences in the document (model test corpus Computer science).

Разброс значений, который виден на небольшом вполне подтверждает идею о том, что статистические данные, собранные на больших корпусах, которые должны насчитывать тысячи текстовых файлов, будет более убедительной. Вместе с тем, поскольку значения соотношения появления в текстах гралогов широкой абстрактной семантики в обоих корпусах находятся в приблизительно одном диапазоне (0,2 – 0,4 в BAWE и 0,2 – 0,3 в model test corpus Computer science) вполне закономерен вывод о том, что эти данные вполне репрезентативны, с ними можно работать, при условии дальнейшего тщательного анализа случаев отклонения от модельных (средних) значений и определения нормы, отклонения от нормы и, тем самым, количественного определения ошибки при дальнейшем применении ПО в учебном процессе.

**Литература**

Фрумкина Р. М., Роль статистических методов в современных лингвистических исследованиях, в кн.: Математическая лингвистика, М., 1973

Орлов Ю. К., Модель частотной структуры лексики, в кн.: Исследования в области вычислительной лингвистики и лингвостатистики, М., 1976

Cai, Y. Instinctive Computing London : Springer. 2016

Wright, D. Stylistics versus statistics : a corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails University of Leeds, 2014

Pogodalla S. Statistics for Corpus Linguistics. URL: <http://www.loria.fr/~pogodall/enseignements/TAL-Nancy/notes-2008-2009.pdf>

Gries, S. Th. Useful statistics for corpus linguistics

URL: http://www.linguistics.ucsb.edu/faculty/stgries/research/UsefulStatsForCorpLing.pd

Gablasova, D.; Brezina, V.; McEnery, T. Exploring Learner Language through Corpora: Comparing and Interpreting Corpus Frequency Information Language Learning, v67 nS1 p130-154 Jun 2017

Special issue: Research report: Data modelling in corpus linguistics: How low may we go? van Velzen, Marjolein H.; Nanetti, Luca; de Deyn, Peter P.. In *Language, Computers and Cognitive Neuroscience*, *Cortex*. June 2014

Weigle, Sara Cushing; Friginal, Eric. Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency In *Assessment in EAP*, *Journal of English for Academic Purposes*. June 2015