

**О статистической состоятельности
диагностических параметров текста**

Г.Я.Мартыненко, Т.Ю.Шерстинова

Санкт-Петербургский государственный университет

gymart@ts4306.spb.edu, tanya@ts4306.spb.edu

1. Среди статистико-классификационных задач компьютерной лингвистики центральной является диагностическая задача. При ее решении исследователи часто используют переменные, параметры которых (средние величины, параметры рассеяния, ассиметрии и т.п.) являются в статистическом смысле несостоятельными, т.е. сильно зависят от объема выборки. Это в значительной степени обесценивает результаты диагностической работы, поскольку филолог с неизбежностью имеет дело с текстами и корпусами разного объема.
2. Проблема сходимости лингвостилистических характеристик текста к предельным величинам, т.е. их состоятельность, рассматривается нами на материале Компьютерной Антологии Русского Рассказа XX века, создаваемой на кафедре математической лингвистики Санкт-Петербургского государственного университета, а также на материале других корпусов, сформированных в России и за ее пределами.

Антология представляет собой полнотекстовую базу данных, включающую около 2500 текстов. Весь корпус “разрезан” на множество хронологических периодов, для каждого из которых может быть построена своя микроантология. В каждую из них включается по возможности максимальное число авторов, писавших в данную эпоху. Для ряда выдающихся писателей строятся свои “авторские антологии” (Чехов, Куприн, Сологуб, Платонов, Зощенко, Горький и др.). На материале всего корпуса в целом, для отдельных хронологических срезов, а также для выдающихся авторов строится система частотных словарей, каждый из которых структурируется с помощью определенной системы параметров.

3. На основании информации, содержащейся в частотном словаре, могут быть построены статистические распределения определенного типа в зависимости от того, какая информация используется в качестве зависимой или независимой переменной.

Основными среди них являются: полиномиальное распределение, спектрное и ранговое распределения. В качестве независимой переменной в таких распределениях выступают: имя лексической единицы (полиномиальное распределение), ее частота (спектрное распределение) и ранг (ранговое распределение), а в качестве зависимой переменной: частота (полиномиальное и ранговое распределение) и число единиц с данной частотой (спектрное распределение).

4. В математической статистике для различных шкал (количественной, порядковой, номинальной) исторически сложились методы обработки данных, пригодные только для работы именно в данной шкале. Самая разработанная система техник сформировалась для количественных переменных. Здесь на основании теории моментов была разработана развитая система средних, показателей вариации, характеристик формы распределения и т.п. Здесь же продуктивно используется система порядковых статистик (мода, медиана, квантили и др.). Что касается обсуждаемых нами распределений (рангового и спектрного), то они, обладая всеми внешними свойствами количественной шкалы, характеризуются патологически большой вариацией признака как на шкале рангов, так и на шкале частот. Это обстоятельство побуждает некоторых исследователей подвергнуть сомнению применимость в данной ситуации теории моментов ввиду их устремленности в бесконечность и попытаться найти какие-то другие характеристики, не реагирующие на объем выборки [1;2]. Что касается полиномиального распределения, то здесь теория моментов не может быть использована в принципе ввиду того, что вариация имеет качественную природу.
5. На основании литературных данных и собственных изысканий нами был сформирован достаточно полный перечень параметров, с помощью которых может быть описана лексикостатистическая структура текста. Каждый из них далее был подвергнут тесту на состоятельность. Все параметры разнесены по группам в соответствии с типом шкалы (см. нижеприведенную таблицу).

Перечень параметров, подвергнутых тесту на состоятельность

Номинальная шкала	Количественная шкала	Порядковая (ранговая) Шкала
Мода (M_o)	*Средняя арифметическая (F_{cp})	Средняя арифметическая (R_{cp})
*Объем словаря (N)	Средняя геометрическая ($F_{г.ср}$)	Коэффициент вариации (V_r)
*Максимальная частота (F_{max})	Коэффициент вариации (V_f)	Медиана ($Me:r$)
	Медиана (Me)	Золотое сечение (Gr)
*Энтропия ($\mathcal{E}$)	Золотое сечение (G)	Среднее линейное отклонение (Dr)

Максимальная энтропия ($\mathcal{E}_{\max}$)	*Коэффициент разнообразия (доля слов с однократным употреблением) (K)	Коэффициент вариации по Dr - $V(Dr)$
Мера упорядоченности ($\mathcal{E}/\mathcal{E}_{\max}$)		Мера концентрации (R_{cp}/N)
		Логарифмический коэффициент концентрации ($\log R_{cp}/\log N$)

*Параметры, упоминаемые в литературе, посвященной решению диагностических задач в лингвистике, социологии, биологии и др. дисциплинах.

6. Проверка перечисленных параметров на состоятельность осуществлялась с помощью методики, основанной на использовании метода наименьших квадратов с рядом принципиальных модификаций, обусловленных сложностью исследуемых зависимостей. Была выдвинута гипотеза, что все параметры сходятся к предельным величинам (альтернативная гипотеза – все параметры возрастают неограниченно). При аппроксимации использовался длинный перечень теоретических функций асимптотического и неасимптотического роста. Наилучшее согласование с эмпирическими данными обнаружила функция Вейбулла (комбинация экспоненциальной и степенной функций).

7. Основные результаты исследования заключаются в следующем:

1) Все параметры имеют теоретический верхний или нижний предел, т.е. являются состоятельными. Однако не все их них можно считать состоятельными практически, так как часть из них достигает верхнего предела лишь при многомиллионных объемах выборки. Однако независимо от объема выборки этот предел может рассматриваться как системная характеристика текста или корпуса текстов.

2) В порядке степени состоятельности пятерка самых состоятельных параметров располагается следующим образом: 1) Мера упорядоченности ($\mathcal{E}/\mathcal{E}_{\max}$); 2) Медиана по рангу ($Me:r$); Логарифмический коэффициент концентрации ($\log R_{cp}/\log N$); Энтропия ($\mathcal{E}$).

3) Предложенная процедура аппроксимации позволяет осуществлять прогноз значений любого параметра при любом объеме выборки.

ЛИТЕРАТУРА

Шрейдер Ю.А., Шаров А.А. Системы и модели. М., Радио и связь, 1982.

Мартыненко Г.Я. Основы стилеметрии. Л., Изд-во ЛГУ, 1988.