# Корпус текстов как семиотическая система и онтология речевой деятельности

# Corpus of Texts – a Semiotic System and Speech Activity Ontology

В.В. Рыков

Московский Физико-Технический Институт
Rykov2000@mail.ru

**Ключевые слова**: корпус текстов, корпусная лингвистика, репрезентативность, фактура речи, онтология, семиотика, речевая деятельность.

Корпус текстов может рассматриваться как достаточно сложно организованная онтология речевой деятельности, отражающую в себе все жанровое разнообразие представленного в нем рода словесности (например - устную, письменную или печатную речь), и занимает промежуточное положение между реальными коммуникативными процессами в обществе, которые он представляет, и формализованной лингвистической теорией, для которой он является источником для исследования. Правильно построенный корпус должен быть организован как формализованная онтология представленного в нем фрагмента речевой деятельности (рода словесности), полученного при помощи определенным образом реализованного процесса так называемой концептуализации. Такой подход к роли и способу составления корпуса противопоставляется так называемому литературоведческому способу, при котором состав текстов корпуса определяется их культурной значимостью.
Также корпус текстов, как сложное словесное единство, включает в себя разнообразную информацию не только о составе и структуре своего речевого материала, но также и другие формализованные методы его представления (индексирование слов, морфологическая информация и т .д.). Следовательно, его также можно рассматривать как специальным образом построенную семиотическую систему. Корпус есть сложно организованное знаковое единство или семиотическая система, денотатами которой являются отраженные в нем различные компоненты речевой деятельности.

|

Термин «онтология» давно уже стал модным в научной литературе. В первом его значении – как описание существенных свойств предметной области - он употреблялся в отечественной лингвистической литературе более двадцати лет назад [2]. Для изучения языка как общественного явления это означало описание реальных коммуникативных процессов, происходящих в обществе [2]. Особенности этого подхода (его условно можно назвать онтологическим) хорошо можно видеть на примере коммуникативных процессов, реализованных при помощи текстов печатного рода словесности, особенно в жанрах художественной литературы. Для того, чтобы коммуникация при помощи печатного текста произошла, он должен быть не только напечатан, поступить в книготорговую сеть, но и прочитан не только критиками, но и массовым читателем. Следовательно, коммуникативный подход означает изучение тех текстов художественной литературы, которые реально читает широкая публика.

Нетрудно видеть, что этот подход принципиально противопоставлен литературоведческому подходу, одним из универсалий которого является констатация дурного вкуса массового читателя. Этот подход сводится к критическому отбору лучших образцов художественной литературы, воспитанию хорошего (по мнению критиков) художественного вкуса у читающей публики [3].

Онтология в описанном выше смысле может рассматриваться как представление в интуитивно понимаемых терминах о предметной области (речевой деятельности) для определенных целей. Тогда составителям корпуса текстов, отражающего даже такой фрагмент речевой деятельности, как художественная литература, нужно прежде всего декларировать, какой подход будет реализован при отборе текстов для корпуса – онтологический или литературоведческий. То есть хотим ли мы видеть в корпусе те тексты, которые читает массовый читатель или то, что им хотелось бы, чтобы он читал. Авторы Брауновского корпуса совершенно ясно декларировали так называемый онтологический подход. Они отбирали тексты для своего корпуса в букинистических магазинах – то, что реально прочитано массовым читателем [5]. Не побоимся назвать это чтивом. Но это реальность и онтология массовых коммуникативных процессов в данной области речевой деятельности. Одним из уникальных свойств такого корпуса является то, что любой его текст может быть заменен равнозначным в смысле процедуры статистического отбора.

Этот подход предопределил успех Брауновского корпуса у самых разных «читателей» и способствовал прогрессу корпусной лингвистики. Ничто не мешает собрать корпус текстов шедевров художественной прозы или поэзии. Только это по существующей терминологии будет называться электронной библиотекой [5].

Далее – если мы теперь будем рассматривать корпус текстов как отражение онтологии речевой деятельности в описанном выше смысле, то тогда и только тогда он будет обладать своими уникальными свойствами. Действительно, лингвистическая теория опирается, как правило, на лингвистические наблюдения или факты, которые, в свою очередь, берутся из речевого материала. Эти лингвистические наблюдения должны быть легко проверяемы, не зависеть от выбранного речевого материала, а также адекватно отражать тот фрагмент речевой деятельности, который стремится описать данная лингвистическая теория.
Таким требованиям отвечает и должен отвечать корпус текстов – особым образом организованное словесное единство. Корпус текстов расположен на машинном носителе, но он отличается от электронного архива или библиотеки. Как уже отмечалось, он также не есть электронное собрание художественных текстов, отобранных квалифицироанными филологами по критерию их культурной значимости [3]. Даже мультимегасловный корпус газетных текстов может отразить только язык газетной публицистики, а не язык в целом и только при условии, что в нем правильно представлены все достаточно разнообразные жанры газетной прозы.

Однако, практика научных исследований показала, что для достаточно сложных предметных областей (таких как, в нашем случае, речевая деятельность) часто необходима структура, занимающая промежуточное положение между представлением о том, что существует в действительности (реальные коммуникативные процессы в обществе) и строго формализованной (в нашем случае линвистической) теорией [4]. Такая структура также называется онтологией, это второе значение этого термина. Такая онтология лежит между тем, что должно быть представлено и его теоретическим обобщением. По-видимому, это вполне соответствует той функции, которую выполняет корпус текстов, являясь с одной стороны достаточно сложно организованной онтологией речевой деятельности и выступая, с другой стороны, в качестве исходного материала для получения новых эмпирических фактов, обогащающих и развивающих лингвистическую теорию.

Следовательно, словесный материал корпуса должен быть организован в онтологическую систему, отражающую в себе все жанровое разнообразие представленного в нем рода словесности (например - устную, письменную или печатную речь). В сущности, правильно построенный корпус должен представлять собой формализованную онтологию представленного в нем фрагмента речевой деятельности (рода словесности), полученного при помощи определенным образом реализованного процесса так называемой концептуализации [1]. Здесь неизбежно приходится обращаться и формализовывать не только состав речевого материала, включенного в состав корпуса, но и к его структуре, а также к другим формализованным методам его представления (индексированию слов, морфологической информации т .д.) – то есть к той компоненте структуры корпуса, которая суммируется одним из четырех классических признаков корпуса – системе его разметки [5].

По сути это специальным образом организованная семиотическая система корпуса текстов. И сам корпус можно рассматривать как сложно организованное знаковое единство или семиотическую систему, денотатами которой являются различные компоненты речевой деятельности.

**Литература**


1. Клещев А.С., Артемьева И.Л. Отношения между онтологиями предметных областей // НТИ. Сер. 2. –М.: 2002. – N 1. – С. 4-23.
2. Котов Р.Г., Якушин Б.В. Онтология языка как общественного явления. – М.: Наука, 1983.
3. Рождественский Ю.В. Принципы современной риторики. – М., 2000.

4. Рыков В.В. Корпус текстов как новый тип словесного единства // Труды Международного семинара Диалог-2003. – М.: Наука, 2003.
5. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh, 1997.

# Corpus of Texts – a Semiotic System and Speech Activity Ontology

**Key words:** text corpus, corpus linguistics, ontology, semiotics, speech activity, knowledge

Text corpus can be treated as a complexly organized speech activity ontology. Really - it describes and represents in itself all the genre variety of real picture and distribution of communication processes in human society – oral, written, printed speech as an area of knowledge. This ontology stands between social speech activity processes which it reflects and at the same time it is the source of data for formal linguistic theory. The corpus as an ontology is the result of specially designed procedures of so called speech activity conceptualization which could be called as processes of sampling and representation as well. This approach stands in opposition to the process of best text selection which is close to the literary criticism and does not reflect the real picture of communication processes in human society. Text corpus as a special kind of word unity includes also various information concerning the genre structure of communication processes it reflects, marking up tokens, word tags etc. Hence text corpus is also a specially organized sign structure or semiotic system. The denotates of this semiotic system are various parts of outer speech activity, its inner properties and organization.