

НАЦИОНАЛЬНЫЙ КОРПУС ЯЗЫКА КАК ИНСТРУМЕНТ ФИЛОЛОГИЧЕСКОЙ ТОПОЛОГИИ

Е.С. Бахтуева (Москва, Россия)

Развитие в конце XX века компьютерных технологий, а вместе с ними и текстовых корпусов, привело к настоящему прорыву в области филологических наук, в том числе в области филологической топологии. В докорпусную эпоху ученые имели возможность проводить топологические исследования лишь на собираемом вручную и поэтому сравнительно ограниченном по объему материале. С появлением электронных корпусов стало возможным выполнение задач, которые в докорпусную эпоху казались неосуществимыми, а именно исследование огромных текстовых массивов устной и письменной речи разных типов и жанров, представленных в национальных корпусах изучаемых языков. Корпус позволяет с применением статистических методов всесторонне и максимально объективно проанализировать языковые и лингвокультурные явления, характерные для данного языкового коллектива на определенном этапе его развития, поэтому основным инструментом филологической топологии на ее современном этапе является национальный корпус языка.

Ключевые слова: *Британский национальный корпус, Национальный корпус русского языка, корпус, национальный корпус, цитата, цитирование, инвариант, филологическая топология*

NATIONAL CORPUS AS A TOOL OF PHILOLOGICAL TOPOLOGY

E. Bakhtueva (Moscow, Russia)

The rapid development of computer technology and electronic corpora in the past two decades has dramatically changed the working procedures in a vast number of language-related studies, including philological topology. In pre-corpus times, topological studies were limited to the analysis of manually collected and, consequently, relatively small text data. National corpora make it possible to cover a large and diverse range of texts and use statistical methods to reveal and carefully analyse linguocultural characteristics of a particular community at a certain stage of its development.

Keywords: *British National Corpus, Russian National Corpus, corpus, corpora, national corpus, quotation, quoting, invariant, philological topology*

Бурное развитие в конце XX века компьютерных технологий, а вместе с ними и текстовых корпусов, привело к настоящему прорыву в области филологических наук, в том числе в области филологической топологии.

Филологическая топология – это быстро развивающееся с конца 70-х годов XX века направление филологических исследований, в фокусе внимания которого находятся проблемы варьирования и установления инвариантности разнообразных филологических объектов – от отдельных слов до целых текстов [Полубиченко, 2017: 3].

Начало этому направлению положили профессора О.С. Ахманова, основатель школы англистики МГУ имени М.В. Ломоносова, и Л.В. Полубиченко, опубликовав в 1979 году в журнале «Вопросы языкознания» статью «Дифференциальная лингвистика» и «филологическая топология», которая провозгласила рождение филологической топологии и обозначила основной круг вопросов данного направления [Ахманова, Полубиченко, 1979].

Задачей филологической топологии является изучение соотношения генетически родственных и/или филологически сопоставимых текстов, выступающих в качестве вариантов, на основе которых постепенно выкристаллизовывается **инвариант**, т.е. абстрактное представление о любом филологическом объекте, полученное путем сопоставления различных вариантов и модификаций данного объекта на основе выделения семиологически релевантных (конституирующих) и отвлечения от несущественных (неконституирующих) параметров, сложившееся в сознании данного языкового коллектива на определенном этапе его развития [Полубиченко, 1991; Бахтуева, 2017]. Иными словами, изучая живой язык путем сопоставления различного рода текстов, филологическая топология призвана представить, что именно является социолингвистически и общественно-исторически важным для определенного языкового коллектива как на современном этапе развития, так и в диахроническом аспекте.

За более чем тридцатилетнюю историю развития филологической топологии можно выделить три основных этапа ее становления как самостоятельного направления филологических исследований [Полубиченко, 2012].

Первый этап (к. 70-х – сер. 80-х гг. XX в.) включает в себя анализ и обобщение всех предпринимавшихся ранее попыток применения топологических идей в лингвистических и литературоведческих исследованиях, а также формирование и обоснование базового понятийно-терминологического аппарата нового направления филологических исследований – филологической топологии.

На втором этапе (сер. 80-х – к. 90-х гг. XX в.) осуществлялись разработка и дальнейшее совершенствование методологии нового направления, в частности методов установления и изучения филологического инварианта.

Третий, современный этап (н. XXI в. – наст. вр.) характеризуется систематизацией огромного потока топологических исследований, которые появляются регулярно и вовлекают в свою сферу совершенно новые материалы.

Основным инструментом филологической топологии на современном этапе ее развития является лингвистический корпус, а точнее национальный корпус языка.

Под **лингвистическим, или языковым, корпусом** текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [Захаров, 2005: 3]. Неотъемлемой частью лингвистического корпуса является **корпусный менеджер** – специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме [Там же].

Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всем многообразии жанров, стилей, территориальных и социальных вариантов и т.п. [Национальный корпус...: эл. ресурс].

Национальный корпус имеет ряд важных, в том числе для топологических исследований, особенностей.

Во-первых, он характеризуется представительностью и сбалансированным составом текстов, т.е. корпус по возможности содержит все типы письменных и устных текстов, представленных в данном языке (художественные, публицистические, научные, деловые, разговорные и т.д.), и все эти тексты входят в корпус пропорционально их доле в языке соответствующего периода [Там же]. Наличие устных текстов в корпусе является бесспорным плюсом для топологических исследований, в частности для топологического изучения цитат, которым и посвящено исследование автора настоящей статьи. Филологи уже давно утвердились во мнении, что изучение живых языков только по письменным текстам без учета их реального звучания является совершенно неправомерным действием: лишь принимая во внимание диалектическое единство письменной и устной форм речи, можно выявить «глубокую взаимозависимость всех трех сторон инварианта цитаты, образующих в совокупности нерасторжимое единство» [Полубиченко, 1991: 127].

Во-вторых, корпус содержит разметку, или аннотацию, т.е. особую дополнительную информацию о свойствах входящих в него текстов [Там же].

В-третьих, языковые единицы представлены в корпусе в естественной языковой среде, т.е. в реально существующих, а не искусственно созданных контекстах, что дает возможность их всестороннего, а значит, максимально объективного изучения.

Более того, корпус позволяет с применением статистических методов сформулировать, подтвердить или опровергнуть некоторую гипотезу о том или ином языковом явлении на большом объеме материала [Корпусная, лингвистика: эл. ресурс].

В докорпусную эпоху ученые имели возможность проводить топологические исследования лишь на собираемом вручную (как правило, при помощи совместных усилий коллег) и поэтому сравнительно ограниченном по объему материале, изучая, например, употребление цитат в художественных произведениях определенных авторов или определенного периода времени. Например, в диссертационной работе В.П. Андросенко, которая положила начало топологическому исследованию цитатной речи, сплошному анализу были подвергнуты цитаты из 90 художественных произведений английских писателей XIX–XX веков общим объемом около 30 тысяч страниц и 45 крупных научных работ ведущих английских филологов послевоенного периода [Андросенко, 1988]. Очевидно, что обследование этих источников и выделение из них цитат требовало немало времени и усилий.

Однако даже на таком сравнительно ограниченном по объему материале удалось сделать ряд важных открытий, которые подготовили почву для других топологических исследований цитат и цитатной речи. В частности, было доказано, что топологические инварианты любых цитат представляют собой диалектическое триединство их структурного (формального), семантического (содержательного) и функционального инвариантов, причем нередко определяющим оказывается лишь какой-то один из них. Впоследствии выяснилось, что данное утверждение справедливо в том числе и для таких генетически родственных с цитатами речений, как поговорки, крылатые выражения, идиомы и т.д. Более того, удалось выявить параллели в эволюции, казалось бы, таких разных по своей сути художественных и научных цитат: подобно тому, как некоторые художественные цитаты постепенно переходят в разряд крылатых выражений, а затем в разряд идиом, некоторые научные цитаты сокращаются и становятся терминологическими выражениями, а затем терминами или терминологическими словосочетаниями [Андросенко, 1988; Полубиченко, Андросенко, 1989].

Только с появлением электронных корпусов стало возможным выполнение задач, которые в докомпьютерную эпоху казались неосуществимыми, а именно исследование огромных текстовых массивов устной и письменной речи разных типов и жанров. Однако широкая доступность языковых данных радикально изменила требования к доказательной базе таких исследований: теперь недостаточно привести несколько примеров в подтверждение выдвигаемой гипотезы, необходимо отобрать и проанализировать значительный по объему корпус фактов языка, подкрепив выводы и обобщения качественного характера статистическими данными.

Впервые национальный корпус, а именно Британский национальный корпус (British National Corpus, далее – BNC), который является общепризнанной информационно-справочной системой и представляет собой собрание в электронной форме всех типов письменных (90%) и устных (10%) текстов середины и конца XX века (общий объем – 100 миллионов слов), был использован в рамках топологических исследований А.П. Корнеевой при изучении функциональных, семантических и структурных особенностей бытования английских фразеологических единиц библейского происхождения в разных типах текстов [Корнеева, 2009]. В общей сложности было проанализировано 486 библеизмов. Данные корпуса позволили не только выявить особенности функционирования библеизмов в речи британцев во второй половине XX века, но и продемонстрировать неуклонное сокращение количества библеизмов, используемых современными носителями английского языка (в активном употреблении находится лишь 9% библейских выражений, еще 19 можно считать среднеупотребительными, тогда как маловостребованными либо вовсе не востребованными оказываются 72% зарегистрированных в словарях библеизмов) [Там же].

Национальные корпуса языков, а именно BNC и Национальный корпус русского языка (далее – НКРЯ), который представляет собой собрание в электронной форме как письменных, так и устных текстов общим объемом более 600 миллионов слов, были задействованы и в нашем исследовании, посвященном онтологии цитат с доминирующим структурным инвариантом и их переводных эквивалентов в современной (сер. XX в. – наст. вр.) английской и русской речи [Бахтуева, 2017]. В данной работе анализировались цитаты, онтология которых определяется прежде всего их формальной стороной. Такие цитаты, обладая четкой запоминающейся формой, часто употребляются в речи не полностью, а в виде рамочной конструкции, слоты которой свободно заполняются любыми значимыми словами, диктуемыми определенной ситуацией.

Исследуемый корпус цитат был составлен на основе самого малого из оксфордских словарей цитат – Little Oxford Dictionary of Quotations (2008, далее – LODQ), который включает в себя около 4000 цитат и представляет собой, по мнению авторов словаря, своего рода антологию заслуживающих внимания речений. Из представленных в словаре цитат была отобрана 2901 предложение из письменных и устных текстов, созданных в XVIII, XIX и XX веках. Из 2901 цитаты, отобранной для исследования, в BNC были обнаружены случаи речевого употребления лишь 515 единиц (18%). Таким образом, анализ контекстов в корпусе показал, что включение того или иного автора в LODQ не всегда является свидетельством его популярности, а зафиксированные в словаре предложения могут и вовсе не цитироваться в современной британской речи. Из 515 цитат 256 речений (49%), встречающихся в корпусе не более одного раза, топологическому анализу не подвергались, поскольку отсутствие в таком случае вариантов использования рассматриваемой цитаты не позволяет сделать вывод о том, что закрепилось в сознании носителей английского языка в виде инварианта соответствующей предложения. В 12 случаях не удалось определить, в каком количестве та или иная цитата представлена в корпусе и/или представлена ли вообще. Таким образом, топологическому анализу были подвергнуты 247 речений, 25 из которых (10%), согласно BNC, являются цитатами с доминирующим структурным инвариантом. Анализ контекстов НКРЯ показал, что переводные эквиваленты 4 из 25 цитат употребляются в русской лингвокультуре преимущественно в виде рамочных конструкций точно так же, как это происходит в британской языковой традиции. Данные корпусов позволили не только изучить особенности реального функционирования цитат с доминирующим структурным инвариантом в современной английской и русской речи, но и подтвердить, что, во-первых, в русской речи также имеет место явление, уже изученное для английской лингвокультурной традиции и заключающееся в употреблении цитат в виде рамочной конструкции, а во-вторых, онтология некоторых цитат такого типа, по существу, идентична в разных языковых традициях.

Таким образом, национальный корпус языка, являясь основным инструментом филологической топологии, позволяет с применением статистических методов всесторонне проанализировать языковые и лингвокультурные явления, характерные для данного языкового коллектива на определенном этапе его развития.

Список литературы

- [1] *Андросенко В.П.* Цитата как элемент сообщения и как фактор эстетического воздействия: дис. ... канд. филол. наук. – М.: МГУ, 1988.
- [2] *Ахманова О.С., Полубиченко Л.В.* «Дифференциальная лингвистика» и «филологическая топология» // Вопросы языкознания. – М.: Наука, 1979.
- [3] *Бахтуева Е.С.* Онтология цитат с доминирующим структурным инвариантом и их переводных эквивалентов в современной английской и русской речи: дипломная работа. – М.: МГУ, 2017.
- [4] *Захаров В.П.* Корпусная лингвистика: учебно-метод. пособие. – СПб., 2005.
- [5] *Корнеева А.П.* Английские фразеологические единицы библейского происхождения в языке и речи: дис. ... канд. филол. наук. – М.: МГУ, 2009.
- [6] Корпусная лингвистика // Фонд знаний «Ломоносов». URL: <http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article>, свободный.
- [7] Национальный корпус русского языка. URL: <http://www.ruscorpora.ru>, свободный.
- [8] *Полубиченко Л.В., Андросенко В.П.* Топология цитат в художественной и научной речи // Научные доклады высшей школы. – М.: Филологические науки, 1989. №3.
- [9] *Полубиченко Л.В.* Филологическая топология: этапы развития и научная проблематика // Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация, 2012. №2.
- [10] *Полубиченко Л.В.* Филологическая топология: теория и практика. – М.: ФЛИНТА: Наука, 2017.
- [11] *Полубиченко Л.В.* Филологическая топология: теория и практика: дис. ... д-ра филол. наук. – М.: МГУ, 1991.