

Зачем и почему мы делаем "Национальный корпус хантыйского языка"?

*Филологический форум Югры – 2024
«Современное филологическое знание в контексте
сохранения и укрепления российских традиционных
ценностей»*

25.09.24

Онина С.В.
доктор филологических наук
профессор Высшей школы гуманитарных наук
Югорский государственный университет

Что такое корпус?

- Электронное собрание текстов,
размеченное таким образом, чтобы в нем можно было быстро найти слова
и конструкции с заданными
грамматическими и другими
интересными лингвисту свойствами.

Что же такое Национальный корпус?

- **Впервые это определение появилось в названии Британского национального корпуса (British National Corpus, BNC), созданного в 1990-е годы в Великобритании специалистами лексикографами.**
- **Национальным корпусом стали называть просто самый большой и представительный корпус, характеризующий язык данной страны в целом.**

Что такое корпус вообще и Национальный корпус в частности

- собрание текстов на данном языке, представленное в электронной форме и снабженное научным аппаратом.
- Собственно, наука о корпусах («корпусная лингвистика») - это прежде всего наука о том, как сделать хорошую разметку корпуса.

Корпусная лингвистика

- В **узком смысле**: как создавать и использовать корпуса текстов
- В более **широком смысле**: как исследовать язык с помощью корпусных инструментов
- Фактически, *новая идеология*

Справка по истории корпусов

- **Первые корпуса : начало 1960-х гг.**

«Брауновский корпус», 1961, 1млн с/у 500 фрагментов по 2 тыс. слов.

- **Современные корпуса: начало 1990-х гг.**

От 100 млн до 1 млрд с/у:

2019, ЮГУ, Ханты-Мансийск

Научный проект

**Лингвистический корпус
югорского регионального текста
как модель сохранения и
взаимодействия аборигенного и
пришлого населения
21.01.2019**



2019, ЮГУ, Ханты-Мансийск

Если возникает жажда гуманитарного знания, пора копать колодцы



- Создать коммерческий образовательный проект как часть образовательного сетевого пространства
- Сформировать на основе контентов и сервисов (в т.ч. на сайте ЮГУ, в Facebook, ВКонтакте и др.) существующих в огромной экосистеме, свой контент для выкладывания корпусов глоссированных текстов, аннотированных исследований (выполненных российскими и зарубежными исследователями совместно с обучающимися – бакалаврами, магистрами, аспирантами, интересующихся тематикой исследования) (по принципу людям нравится что-то выкладывать в сеть и показывать другим)
- Образовать в открытом доступе образовательные проекты прикладного плана понятных для потенциальной аудитории, издавать и выкладывать лекции ведущих специалистов мирового уровня Финляндии, Венгрии, Германии, Австрии, России и др. в области исследования обско-угорских языков, таким образом монетизируя проект (при грамотном подходе инвесторам легче найти стартовый капитал или спонсоров и что они могут быстрее монетизироваться (с точки зрения организации внутренних процессов и программной реализации они решают в основном те же задачи).

Корпуса сегодня: параллельные корпуса

- Корпуса с языковыми парами (25 языков, от английского до эстонского) и многоязычный корпус (ок. 20 языков, несколько текстов)
- Наиболее крупные:
- Английский (44 млн), немецкий (30 млн);
- Небольшие:
- Китайский, латышский, чешский, финский (по 4 млн), армянский (1,5 млн) ...

НКРЯ сегодня: специальные корпуса

- Диалектный (600 тыс.) образцы
- Обучающий (600 тыс.) школьная программа
- Синтаксический (1,5 млн)
- МультиПарк (600 тыс.)
- Акцентологический (13млн): устный + поэтический + корпус наивной поэзии, история русского ударения

Как быть? Что искать?

Разметка корпуса: если мы хотим искать не слова, а формы

- Представьте себе такую задачу: найти в достаточно длинном тексте все формы дательного падежа единственного числа.
- Никакой текстовый редактор с этой задачей, конечно, не справится.
- На самом деле хорошо размеченный текст для специалиста оказывается **совершенно бесценным.**

Размеченные корпуса

- Первый серьезный инструмент, позволяющий существенно ускорить и упростить эту процедуру.
- Разумеется, для этого надо сначала сделать корпус. Но **хороший корпус делается для многих**, так что мы надеемся, что **наш труд будет востребован**.

Объём национального корпуса

- Объем измеряется сотнями миллионов словоупотреблений
- Но, кроме того, он - и это даже важнее - должен быть представительным.
- Иначе говоря, он должен содержать все типы текстов, представленные в данном языке в данный исторический период.

Национальные корпуса сегодня (НКРЯ)

- Один из первых корпусов, созданных лингвистами и для лингвистов - **НКРЯ**
- Развивается с апреля 2004 г.
- Более 2 млрд с/у, весь период существования русского языка (XII – XXI вв.)
- Охватывает все виды текстов
- Имеет метатекстовую, грамматическую, семантическую и специальную аннотации

Что такое "Национальный корпус хантыйского языка" в перспективе?

- **Национальный корпус хантыйского языка (НКХЯ, Корпус) — в перспективе - это собрание независимых корпусов, каждый из которых предназначен для решения определенных лингвистических задач.**
- **Каждая из этих коллекций текстов в перспективе должна стать большой по объёму и представительной.**

В перспективе НКХЯ

- Национальный корпус хантыйского языка не должен ограничиваться, например, только произведениями художественной литературы / поэзией (желательно на хантыйском языке), сколь бы важны они ни были для изучения хантыйского языка.
- Он будет содержать газетные статьи разной тематики (от общественно-политических до, например, спортивных), в идеале ... и специальные тексты
- в Корпус попадают образцы практически любого существующего в хантыйском языке письменного дискурса

В перспективе НКХЯ

- Более того, для полного и адекватного представления о том, что происходит в современном хантыйском языке, необходимо расширить рамки Корпуса и включить в него и записи устной речи

Источник устных текстов

- Записи радиопередач, интервью
- Но, не ограничиваясь этим, также записи бытовых разговоров (диалогов и монологов) в домашней обстановке.
- В Корпусе, конечно, должны быть образцы настоящей диалектной речи.
- Тексты так называемой «электронной коммуникации»

Кому и зачем может быть нужен такой корпус?

- Для лингвистов
- фактически справочно-информационная система по современному хантыйскому языку, позволяющая получать ответы и
- позволяющая ставить новые проблемы

Кому и зачем может быть нужен такой корпус?

- Корпус технически сокращает и революционизирует работу с материалом
- В перспективе выполнение студентами практических заданий, курсовых и дипломных работ
- В перспективе на материале Корпуса будут защитаться диссертации.
- Корпус, словарь и грамматика, скорее всего, соединятся в единый электронный ресурс - или базу данных, на основании которой и можно будет изучать хантыйский язык.

Кому и зачем может быть нужен такой корпус?

- Специалистам – программистам
- Поскольку программы такого рода имеют дело с естественным языком
- Причем - внимание! - понимать структуру именно тех текстов, которые встречаются в реальной жизни

Кому и зачем может быть нужен такой корпус?

- Не случайно проект создания Национального корпуса мансийского языка уже на самых ранних этапах был поддержан компанией «Яндекс»
- В целом вне лингвистики наибольшее внимание к Корпусу проявляли и проявляют специалисты по информатике и программированию.

Кому и зачем может быть нужен такой корпус?

- Преподаватели современного хантыйского языка, скорее в качестве иностранного;
- В российской школе Корпус может быть полезен для учителей и учеников;
- Ничто не мешает учителю и его ученикам составить собственный подкорпус

Кому и зачем может быть нужен такой корпус?

- Люди пишущие и, шире, вообще как-то в своей повседневной деятельности связанные со словом.
- Например, **редакторы газет и журналов**, которым приходится решать вопросы узуса:
- Допустимо ли такое слово или конструкция?
- Кто, где, когда употребил впервые такой оборот?
- Для каких типов текста он наиболее характерен?

Трудно представить себе, где можно быстро найти ответы на такие вопросы, если не в Корпусе.

Кому и зачем может быть нужен такой корпус?

- Корпус дает для изучения хантыйского языка и понимания закономерностей его развития, ...
- **Проследить за хронологией постепенных языковых изменений** (ср.: типа Дв.ч.).

Что же нам может сказать Корпус
о современном состоянии хантыйского языка?

- С точки зрения истории хантыйского языка, нынешний период является чрезвычайно важным и интересным.
- Это - период языкового сдвига, может быть - языкового слома.

Что же нам может сказать Корпус о современном состоянии хантыйского языка?

- Изменения лексического состава: (ср.: *верты* 'делать' – *вератты* 'мастерить');
- Изменение произношения (в речи современных пользователей не различаются звуки *л* - *л̣* : *вўлы* вместо *вўлы* 'олень'; *н* – *ң*: *пўхнән* вместо *пўхңән* и т.д.)

Корпус будет!!?

- *Ждет ли хантыйский язык судьба языков, потерявших все падежи, функции которых взяли на себя предлоги, или вообще исчезнет язык - конечно, ни один серьезный специалист предсказывать не берется. Но «что-то может произойти».*
- **Теперь** у нас, надеюсь, **будет Корпус**.

Сора кеша!
Тăлаң йош, тăлаң кӳр!

Спасибо!
Будьте здоровы!