Конструирование корпуса письменных текстов учеников болгарской средней школы. Вопросы. Трудности. Решения.

Towards Building a Corpus of Students' Written Texts from Bulgarian Secondary School. Problems. Difficulties. Decisions

д-р Татьяна Ангелова

(Софийский университет им. Св. Кл. Охридского, Болгария)

0. Почему необходим корпус речи учеников? Что вызывает необходимость создания корпуса текстов учеников на болгарском языке?

Перед современными исследователями, выбравшими заниматься методикой обучения болгарскому языку, возникают вопросы, среди которых можно особо выделить проблемы, связанные с изучением продвижения и постижений учеников в процессе обучения болгарскому языку, с результативностью использования альтернативных учебников, с усовершенствованием коммуникативно-речевой и социально-коммуникативной компетентности учеников, с оценкой и измерением не только знаний по болгарскому языку учеников, но и их умений использования текстов в определенном социокультурном контексте, общаться преемлимо, адекватно. Не на последном месте нужно учитывать ошибки и недостатки, типичные для речи учеников и их текстов (всех текстов, созданных в ходе обучения болгарскому языку обучаемыми) и необходимость систематизировать их в общей национальной базе данных, которая будет находиться в распоряжении учителей и методистов, тестологов. Здесь мы только намечаем вопросы, без стремления полностью их исчерпать.

Адекватным исследовательским инструментом для решения этих вопросов оказываются процедуры корпусной лингвистики, сбор и обработка базы данных из корпусов текстов, отражающих речевое поведение при общении в разных социокультурных сферах.

Не является дискуссионной необходимость в изучениии речи учеников, в изучении речевого поведения учеников в формальной обстановке общения – обучения болгарскому языку, исследовать влияние неформального общения вне занятий, вне школы на коммуникативно-речевую компетентность учеников. Предметом обсуждения, однако, являются разные виды деятельности, с помощью которых можно *верифицировать* методические исследвания, чтобы они залегли в основу государственных образовательных требований по болгарскому языку, учебников, обучающих деятельностей, деятельностей по оценке обучению болгарскому языку, оценке достижений учеников по болгарскому языку, даже результативности работы учителя по болгарскому языку; изучать и улучшать качество обучения и образования по болгарскому языку.

Сбор текстов учеников позволяет накопить богатый эмпирический материал о продвижении вперед учеников в их языковом развитии, об их подготовке по болгарскому языку (как процессе и результате обучения болгарскому языке), о типичных недостатках и ошибках в употреблении языковых единиц, в словоупотреблении, в создании определеных типов и видов текста.

Во многих местах собраны текстовые архивы региональных инспекторатов, отдельных школ. Но эти тексты очень редко используются, если не сказать, что они *никак* не используются в исследовательских целях, а даже и в практических целях. Их основное предназначение – отметить в отчете очередную деятельность и удовлетворить конъюнктурные задачи. В нашем изложении мы не будем ставить административных проблем, связанных с конструированием корпуса текстов учеников, но попытаемся определить, какие аспекты этого процесса методически и методологически значимы и что возможно реализировать в современном образовательном контексте болгарской школы по отношению к этой проблеме.

Вот почему мы остановимся на нескольких узловых вопросах:

- Что представлаяет собой корпус текстов учеников?
- Через какие этапы проходит его дизайн?
- Какие трудности возникают? Как можно использовать накопленный опыт (опыт WordNet, Лингвистического консорциума и других источников)? Каковы пути создания корпуса текстов учеников возможны: от эмпирического материала к модели, от модели к эмпирическому материалу?
- Какие задачи можно решить с помощью корпуса текстов учеников на болгарском языке (КТУБЯ)? Каково его приложение?

1. Что представляет собой корпус текстов учеников? Корпус текстов учеников - компонент Национального корпуса болгарского языка, и соответственно отвечает требованиям к корпусам текстов.

Вот какая дефиниция о корпусе предложена в Словаре WorldNet (в интернете). http://www.hyperdictionary.com/

WorldNet Dictionary

Definition:

1. [n] the main part of an organ or other bodily structure
2. [n] a collection of writings; "he edited the Hemingway corpus"
3. [n] capital as contrasted with the income derived from it

Synonyms: principal, principal sum

See Also: accumulation, aggregation, assemblage, body part, capital, collection, corpus striatum, part, piece, striate body

Корпус ни в коем случае не механический сбор текстов. Как подчеркивают Тони Макенери и Андрю Уилсон, корпус может состоять хотя бы из двух текстов. Больше чем один текст уже образует корпус (срв. латинское слово *корпус* означает на самом деле тело (текста). Но если рассмотреть его в контексте лингвистических исследваний, текстовой корпус должен отвечать определенным требованиям: [McEnery T., A. Wilson 2001: 21] .

Представительность (Sampling and representativeness) - первое требование к корпусу из текстов учеников. Это означает, что он должен быть представлен через *примеры, образцы,* апробированные и типичные, существенные для корпуса, т.е. они являются *представительными для корпуса*. С лингвистической точки зрения все языковые варианты составляют интерес для исследователей. В таком смысле имеет значение какому способу сбора данных языковых вариантых отдается предпочтение. Есть два способа сбора данных: а) можно проанализировать каждое высказывание в отдельности в этом многообразии, и б) можно сконструировать небольшой образец вариативности высказываний, который является представительным. Стандартизирование корпуса достигается чаще свего вторым способом. Важно отметить для корпуса текстов учеников, что он должен включать автентичные тексты, т.е. тексты, созданные учеником без вмешательства учителя, взрослого. В нашем случае мы имеем в виду тексты учеников, создаваемые в рамках занятий по болгарскому языку и литературе, а не в домашних условиях, когда ученик может списать, написать сочинение вместе с родителями и т.д. Конечно, было бы очень полезно собрать тексты учеников с занятий по истории, географии, другим учебным предметам. Но на этом этапе такое ожидание звучит скорее утопически. Представляется осуществимой задача сбора и обработки письменных текстов учеников. Устные тексты учеников (как и тексты, создаваемые во время педагогической коммуникации в болгарской средней школе) также представляют собой объектом исследовательского интереса, но из-за их специфики они рассматриваются как объект после письменных.

О представительности корпуса можно судить и в зависимости от задач, которые он обслуживает. Например, если исследователь хочет собрать корпус из аргументативных текстов учеников 5-ого – 7-ого классов, очевидно нужно собрать сочинения, в которых преобладают рассуждения, тезис играет конституирующую роль. Представительность охватывает как количественные, так и качественные показатели.

Второе требование – порог количества, ограниченный объем (Finite size) относится к *числу единиц в корпусе*. Практика показывает, что количество единиц, содержащихся в корпусе – 1 000 000 слов, например, - ориентировочный объем. Этот тип корпусов называется закрытым. Раз собранный, эти данные не расширяются, а остаются в распоряжении исследователей. Есть и т.наз. называемые открытые корпуса. К ним можно непрерывно добавлять новые данные, и такой вид корпуса называют *мониторным корпусом* (Дж. Синклер). Второй вид корпуса позволяет избежать устарение введенных данных, а первый вид корпуса рассчитывает на надежность в количественном отношении собранной информации. Закрытый тип корпуса предпочитается при обработке архивов состязательных экзаменов, потому что собранное число текстов - крайний, ограниченный состязательными условиями. Открытый тип корпуса лучше обслуживает ежедневную практику обучения, нужды учителя, школы, региона (в зависимости от социокультурных условий обучения, типичный пример – это тексты из двуязычной среды обучения).

Третье требование– это *цифровизация* (Machine-readable form) – т.е. тексты могут быть прочитаны машиной (компьютером). Специалисты называют этот процесс *дигитализацию текстов*. Дигитализация позволяет легко обеспечить тексты корпуса лингвистической информацией, или другими словами быть аннотированными. Шифрование и аннотирование текстов требует специального внимания. Здесь мы ограничимся лишь замечанием, что шифрование делается с помощью т.наз. *тэгов.* Например, указывается, что слово является существительным именем, мужского рода единственного числа, и это отмечается специальным означением, понимаемого машиной. Эти единицы называются *справочными единицами* (Entity reference).

Аннотирование может осуществляться на разных уровнях (от фонетического до текстового) и с помощью разных схем (т.наз. схем и языков для маркирования – SGML, XML и др.).

Четвертое требование (A standard reference) к корпусу заключается в том, что он должен служить больше общепризнанным справочником, который "соответствует" использованной методологии. Он скорее всего должен быть адекватным, корректным по отношению к определеной методологии, исследовательским нуждам, чем предствалять все оттенки, различия в языковых вариантах. Иными словами, исследователь должен ссылаться на примеры из корпуса. Если привлечь тот же пример об исследовании умений учеников создавать аргументативный текст, в корпусе следовало бы разместить образцы, адекватные предпочитаемой методологии – прагматической, текстолингвистической, реторической и т.д.

2. Через какие этапы проходит дизайн корпуса текстов учеников? Для нужд изложения мы использовали справочную информацию с рекомендательным характером лингвистического консорциума о базе данных (Linguistic Data Consortium). Она включает следующие ключевые моменты:

А. На первом месте указывается методология деятельности по сбору текстов, происхождение собранных текстов (Background).

Под происхождением текстов понимается информация о том, где они собраны, во время занятий по болгарскому языку, а не в домашних условиях, т.е. стремление, чтобы собранные данные были автентичными, представляли тексты учеников, а не обучающих их учителей. Важно отметить при аннотировании текстов, что они являются спонтанной речью, созданы в процессе написания, а не редактированы под влиянием обучающего. Возможно включить тексты - результат обучающих усилий учителя. Условно такие тексты могут быть названными неспонтанными. Строго говоря, письменная речь всегда имеет характер подготвленности, обдумывания, письменный текст редактируется, чтобы он воспринимался лучше. Тем более, что письменное общение чаще всего официальное общение, стремление автора представить себя в положительном свете. Нужно отметить, что это не всегда осознанный мотив для пишущего ученика.

Возможно при сборе текстов исходить из определенной методологии, т.е. иметь в виду определенную теорию. Работающие постановки, например, - это понимание о сути речевого поведения, речи, коммуникации. Популярны теории речевой деятельности, речевых актов и т.д. Независимо от того, какие теоретические постановки используются, в конечном счете важно достичь согласия между исследователями по поводу: задач, для которых будет использован корпус, текстов, которые будут собраны, их количества и т.д. Само собой разумеется, что подобная деятельность, как создание корпуса текстов учеников, можно выполнить только коллективно, а не индивидуально. Например, целенаправленно собирать только аргументативные тексты, ученков определенного возраста, с определенной социокультурной характеристикой – из большого города, маленького города, деревни; от профилированной / непрофилированной школы; указать на то, были ли отобраны ученики – авторы текстов в результате конкурсного экзамена или нет. Будет ли "открытым" корпус или будет содержать определенное число текстов, и на этом корпус будет финализирован.

Б. Сбор базы данных (Data Collection) занимает второе место. Прежде чем начать произвольно собирать базу данных, целесообразно найти ответ на вопрос, в какой степени можно использовать личные архивы, архивы региональных инспекторатов, олимпиад по болгарскому языку и литературе, архивы вступительных экзаменов. Кто будет собирать тексты и вводить в компьютер. Как это будет просиходить? Возможно сохранить их автентичный вид при помощи скенирования. Наряду с этим тексты можно набрать, чтобы их

восприятие и обработка были удобны. С методической и прагматической точки зрения функционально сочетать оба подхода.

В. Третья деятельность – аннотирование базы данных (Data Annotation). Иными словами, это снабжение текстов корпуса лингвистической информацией – примечания, комментарии. В специализированной литературе (см. указ. источник) представлены определенные схемы и типы аннотирования. Решается, какие виды аннотирования использовать: заглавие текста и сам текст. Нужно также уточнить:

- текстовую и внетекстовую информацию – например, в заглавии также отмечается тема текста, возраст ученика, девочка / мальчик; дата создания текста; урок болгарского языка или урок литературы, в какой школе, каком классе и т.д. Таким образом, можно искать среди множества файлов по определнным параметрам. Кто записал текст, кто его проверял?
- орфографию или способ транскрибирования. Напр., Формат *Какао* (Cocoa)
- лингвистическую аннотацию через морфологическую разметку по частям речи (тэгирование) и лемматизацию (все варианты оформляют лемму лексемы (понимаемой как объединение трех основ слова – словоизменительной, формообразовательной и словообразовательной), напр., *удари, ударен, удряне 'удары, ударный, ударять* – отглагольное существительное' оформляют лемму *удар'удар'*; срв. понимание леммы как основной формы слова, которая имеет множество словоформ в зависимости от выраженных грамматических форм – см. http://www.larflast.bas.bg/balric/index/index.htm).
- парсинг – парсирование, будучи морфосинтаксической категорией, представляется с помощью т.наз. "синтаксических деревьев".
- семантику или семантические особенности слов, их значения и семантические отношения между словами.
- дискурсную и текстолингвистическую информацию – напр., прагматические маркеры, анафорическую аннотацию (когезию текста; конкретно, напр., коллокационную когезию). Основной вопрос, который возникает здесь, как аннотировать ошибки учеников (текстовые, структурные, конвенциональные и т.д.)
- фонетическую транскрипцию, которая касается только звуков, пауз и т.д.

Кодирование тесно связано с аннотированием. Оно представляет собой использование определенного знакового языка для представления информации из корпуса.

Г. Четвертая деятельность - это парадигматическое построение базы данных или систематизированной базы данных, их стандартизирование в образцах. Построение базы парадигматических данных становится возможным с помощью кодирования и аннотирования (Building Paradigmatic Data).

Д. На предпоследнем месте идет управление корпуса - кто отвечает за информационную поддрержку, как и где сохраняются тексты. Нужно уточнить институциональные и финансовые требования к сохранению и поддержке корпуса (Management of Corpus Building Effort).

Е. Не на последнем месте по значению приложение и распространение базы данных из корпуса текстов учеников (Data Dissemination). Предпочтительно, чтобы это было осуществлено институционально. Практика показывает, что используются возможности финансирования или со стороны определенного высшего учебного заведения, решающее исследовательские задачи, или со стороны Министерства образования и наук Республики Болгарии. Трудности первого выбора относятся к финансовой стороне, материальной базе – компьютеры, информационный портал, который может поддерживаться для корпуса ученической речи. Нельзя пренебрегать и решение, при котором работа над корпусом присоединяется к уже существующим проектам, доказавшим, что они хорошо функционирующие. Риск при втором выборе заключается в тяжелой бюрократической процедуре.

Плюс – в институциональном обеспечении. В любом случае важно обеспечить взаимовыгодное сотрудничество между странами, заинтересованными в создании, поддерживании и использовании корпуса.

3. Какие трудности возникают? Прежде всего нужно решить вопрос о пути, по которому будет создан корпус текстов учеников. Возможные пути к созданию - от эмпирического материала к корпусу, от модели к эмпирическому материалу. И один, и другой подход имеют преимущества и недостатки. Возможно их сочетание.

Трудности можно условно обособить в несколько групп: *методологические* (какая программа будет обслуживаться, какова теоретическая платформа), какие специалисты будут объединены общей целью: методисты, лингвисты, программисты, статистики, психометрики, учителя и т.п. Не менее существенные трудности порождаются квалификацией участников, нуждой быть обученными в областти корпусной лингвистики и, разумеется, чтобы они были специалистами в собственной научной области. Эти трудности условно можно определить как *кадровые* и *управленческие*. Нужно ответить и на вопрос, кто будет ”печатать”, вводить тексты на компьютере. Этой деятельностью могут заняться: студенты, дипломники, магистранты, молодые и мотивированные учителя, которые имеют компьютерную грамотность.

На третьем месте трудностите, имеющие отношение к материальной базе – чтобы в распоряжении было хотя бы два компьютера, сервер, ксерокс, сканер, и т.д. Должен быть изготовлен бюджет по созданию и поддержке подобного звена. В самом общем виде это *финансовые* трудности.

Для преодоления некоторых из этих трудностей можно использовать накопленный опыт (WordNet, Linguistic Data Consortium). Некоторые из трудностей могут оказаться непреодолимыми.

Нам хочется отметить, что самый убедительный аргумент для преодоления возможных трудностей остается образовательный и научно-исследовательский потенциал такого корпуса.

3.Какие задачи могут быть решены с помощью корпуса? Каково его приложение? Трудно сказать, какие задачи – научно-исследовательские или образовательные – более важные. Они тесным образом связаны. С помощью корпуса текстов учеников можно создать базу данных о постижениях и трудностях учеников в усвоении определенного вида умений, определенного вида учебных жанров, определенной обратной информации по осуществлению целей обучения болгарскому языку. Создается объективая картина о том, как соблюдаются, “покрываются” государственные образовательные требования по болгарскому языку в соответствии с накопленным и обработан эмпирический материал. Интерпретация данных становится в соответствии с решаемыми задачами: методическими, методологическими, узко практическими и т.д. С помощью данных корпуса можно будет составлять корректные тестовые задачи и подготовлять дидактические тесты не только для нужд школьная практика по болгарскому языку, но и для стандартизированных тестов, обслужвающих национальные экзамены, лингвистические олимпиады, диагностические исследования языкового и интеллектуального развития учеников в средней школе.

Литература:

1) Лингвистический консорциум для базы данных Linguistic Data Consortium http://www.ldc.upenn.edu/
2) **McEnery** T. and **A. Wilson** Corpus Linguistics. An Introduction. Edinburg University Press. 2001
3) http://www.larflast.bas.bg/balric/index/index.htm
4) Corpus Linguistics and Language Teaching www.listserv.linguistlist.org/archives/cllt.html; http://www.linguistlist.org/list-archives.html; www.ruf.rice.edu/~barlow/cllt.html;
5) Красимира Петрова Проект о создании корпуса устной речи русско-болгарских билингвов. Rethrieved from www.dialog-21.ru 2003