Проект о создании корпуса устной речи русско-болгарских билингвов

Красимира Петрова

В последние несколько лет болгарские исследователи и научные работники объединяют свои усилия для создания **национального корпуса болгарского языка**. Отдельные его части разрабатываются в Институте болгарского языка (http://www.ibl.bas.bg) и Лаборатории по лингвистическому моделированию (http://www.lml.bas.bg) при Болгарской академии наук, в Софийском университете им. Св. Климента Охридского и в других вузах страны, а также в некоторых фирмах, разрабатывающих программы и средства для компьютерной обработки естественного языка (http://www.sirma.bg/ontolex; http://www.bultra.com; http://tran.skycode.com), международной группой, работающей над компьютерной обработкой средневековых славянских рукописей и ранних печатных книг (http://clover.slavic.pitt.edu). Встает вопрос о стандартизации и унификации этих серьезных, но разрозненных усилий, чтобы сделать эти ресурсы сопоставимыми и совместимыми с воспринятыми мировыми стандартами и обеспечить возможность их обработки существующими программами. В рамках магистрских программ по лингвистике (http://www.slav.uni-sofia.bg/Pages/LING.HTM), по компьютерной гуманитаристике (http://www.slav.uni-sofia.bg/Pages/COMHU.HTM) на Факультете славянских филологий в Софийском университете, на спецкурсах по корпусной лингвистике, а также и благодаря энтузиазму отдельных исследователей – создателей первых электронных ресурсов по разговорной речи болгарского языка (http://www.hf.uio.no/easteur-orient/bulg/mat/) делается попытка создать подкорпус болгарской разговорной речи.

Нам кажется интересным и полезным создать раздел в **подкорпусе** по болгарской речи, продуцируемой не-носителями болгарского языка, а также раздел корпуса, посвященного устной / письменной речи **билингвов / мультилингвов**, для которых одним из языков является болгарский. На отделении русской филологии собирается аудиоматериал по устной разговорной речи билингвов с первым языком как болгарский, так и русский. Этот материал – полезный источник для иследований в области контрастивной, психо-, социо-, этнолингвистики. Уже появляются первые курсовые и дипломные работы в этом направлении. Исследуемые проблемы – интерференция на всех языковых уровнях, переключение кода и др.

В создании корпуса должны учитываться различия между **архивом, корпусом и базой данных** – корпус занимает промежуточное положение между архивом и базой данных [см. Синклер 2000].

Согласно требованиям современной корпусной и компьютерной лингвистики, дизайн корпуса должен соответствовать определенным требованиям в зависимости от его предназначения и целей [см. Гюли, Пиперидис 2003 - материалы по третьей теме международного проекта по созданию Балканских региональных информационных центров по повышению осведомленности и стандартизации лингвистических ресурсов и программ и передовых технологий в области обработки естественного языка - http://www.larflast.bas.bg/balric/]. Минимальные требования, чтобы данный текстовой архив можно было считать корпусом, следующие [там же]:

- **формат и сохранение** в читаемой машиной электронной форме.

Чтобы использовать уже созданные средства для компьютерной обработки подобного ресурса, текст должен быть затранскрибирован максимально близко к орфографическим нормам кодифицированной письменной речи. Для распознавания отдельной словормы морфологическим анализатором для болгарского языка Slovnik, который распознает более 110 000 болгарских лексем (http://www.diogenes.bg/slovnik/index.html (существует еще несколько морфологических словарей http://www.pu.acad.bg/dcs/lingua1.htm и морфологических анализаторов - таггеров для болгарского языка – см. демо-версию, основанную на системе MORPHO-ASSISTANT http://www.larflast.bas.bg/balric/bulric_home.htm). Все слова, отличающиеся по графике от зафиксированных словоформ литературного языка, а также слова – вкрапления из “другого” языка,

должны быть представлены отдельным списком – это одно из возможных решений дальнейшей автоматической обработки текстов. Транскрибированные тексты должны быть конвертированы согласно стандартам TEI (Text Encoding Initiative – http://tei-c.org/, http://www-tei.uic.edu/orgs/tei , совместимые с XML , чтобы потом можно было обрабатывать программой ClaRK, основанной на формальной грамматике HPSG для создания болгарского банка синтаксических деревьев (см. http://www.bultreebank.org). Гибкая система позволяет аннотирующему вводить собственные знаки маркирования, которые представляют интерес для исследователя.

Другая возможность машинной обработки подкорпуса разговорной речи русско-болгарских билингвов – это программы TRACTOR и TRASA, которые были адаптированы для работы с кириллическими текстами и выработана система транскрипции [см. Алексова, Петрова 2002], которая "читается" программами, в совместном проекте – семинаре между Университетом в Гетеборг и Софийским университетом (http://www.ling.gu.se/~leifg/sofia/; http://www.qualitative-research.net/fqs-texte/3-00/3-00allwoodetal-e.pdf).

- **структура -** необходима, чтобы можно было делать выводы на основе наблюдений согласно первоначально задуманным критериям дизайна корпуса. В проектируемом корпусе русско-болгарских билингвов должны быть четко выделена часть, описывающая внешнюю информацию (header - описание записанных лиц, родной, второй и т.д. языки, срок проживания в иноязычной среде или срок изучения иностранного языка, пол, возраст, профессия и др. - все факторы, значимые для психо-, социо-, этно- и контрастивных лингвистических исследований) и сам текст (body). Особенно важно отражать диалогическую структуру текста и участие коммуникантов.
- **размер -** должен быть "достаточным" в зависимости от поставленных целей; нет строгих границ. В нашем случае зависит от наличия четких ясных аудио-, а впоследствии и видеозаписей общения русско-болгарских билингвов.
- **представительность -** в зависимости от целей корпус должен быть преставительным для разнообразия языков и подъязыков, которые он должен отражать. В наших целях необходимо записать билингвы с родным языком русским и болгарским. Они должны быть представителями разных социальных и возрастных групп.
- **баланс -** корпус пропорционально должен охватывать разные типы текстов в рамках языковой общности. Мы планируем записи разнообразных сфер человеческой деятельности и общения - профессиональное, бытовое, неформальное и т.д. общение.

Соблюдая перечисленные требования, мы получим стандартный корпус многократного использваия для решения поставленных лингвистических и методических задач. Согласно спецификации проекта EAGLES (http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html) для корпусной типологии, задуманный нами корпус будет **подкорпусом** национального корпуса болгарского языка, он должен подчиняться выбранным характеристикам и параметрам корпуса, к которому принадлежит, и будет отражать **подъязык** русско-болгарских билингвов разного типа, проживающих в Болгарии. Придерживаясь классификации корпусов, предложенной проектом EAGLES, наш корпус будет отличаться следующими характеристиками:

- **по модусу, форме существования** - транскрибированные тексты в письменной форме с постепенным накоплением дигитализированных аудио- и видеозаписей. Лучшая перспектива - исследование мультимедийного материала с помощью программ TRACTOR и TRASA, разработанных в университете в Гетеборг и адаптированных для болгарского языка (http://www.ling.gu.se/~leifg/sofia/).
- **по типу текста** - устный и затранскрибированный корпус;
- **по объему** - небольшой подкорпус;
- **по языковому покрытию** - корпус подъязыка; он должен быть достаточно представительным, чтобы можно было делать значимые статистические обработки для

достоверных наблюдений и заключений (конкретные данные подлежат дальнейшей дискуссии).

- **по жанру (регистру)** - записи разговорной речи (при наличии литературных, научных, юридических, технических и т.д. текстов в целом корпусе).
- **по варьированию языков** - задуманный нами подкорпус не будет ни одно-, ни многоязычным, ни параллельным переводным в традиционном понимании, а его своеобразие как раз заключается в "смешении" русского и болгарского языков, что является объектом нашего наблюдения.
- **по общности, которая продуцирует корпус** - охватывает билингвов - носителей русского и болгарского языков.
- **по маркированности** - на первых порах готовится коллекция "простых, чистых" текстов, а уже потом можно думать об их таггировании и аннотировании.
- **открытость корпуса -** необходима возможность постоянного пополнения и накопления данного подкорпуса.
- **по национальному варьированию -** это часть корпуса болгарского языка, с возможной перспективой исследования речи не-носителей болгарского языка, говорящих на болгарском языке (например, есть идея использовать материалы телепередачи "Далекоглед" (`дальнозоркая труба`), которая уже несколько лет передает интервью с иностранцами, владеюшими болгарским языком в разной степени и пребывающими в течение разного срока в Болгарии); есть намерения также исследовать язык болгар, проживающих длительное время за границей, а также включить и другие комбинации языков билингвов / мультилингвов.
- **по историческому варьированию -** это синхронный, а не диахронный подкорпус;
- **по географическому / диалектному варьированию -** на первом этапе записи проводятся только на территории Болгарии;
- **по возрасту -** нас интересует речь как детей, так и взрослых;
- **по доступности, наличности подкорпуса -** мы хотим сделать этот корпус свободно доступным в Интернете для изучения и использования в научных целях; это будет частью национального корпуса болгарского языка.

Почему нам кажется интересным и необходимым изучать речь русско-болгарских билингвов?

Болгария из тех стран, где используется преимущественно один язык для общения. Тем не менее, есть группы населения, пользующиеся несколькими языками. Такова группа русскоговорящих (носителей русского языка), постоянно проживающих в Болгарии. Трудно собрать официальные данные об их количестве, но во время последней переписи населения в Болгарии в 2001 году по данным национального статистического института около 15 000 русского происхождения (данные предоставлены ассистентом Кафедры русского языка Софийского университета А.С.Барановой, работающей над проблемами русско-болгарского билингвизма). Реальные цифры русско-болгарских билингвов с родным языком русским намного превышает указанные, так как к ним нужно отнести и их дети, а также множество представителей других бывших Советских республик, говорящих на русском языке.

Объект исследования, для которого создается корпус, - идиолекты русско-болгарских билингвов (с родным языком как русский, так и болгарский), а предмет – явления, характеризующие устную речь билигвов – интерференция на всех языковых уровнях, “переключение кода”, особенности речевого этикета, речевые стратегии, коммуникативные техники и др., в зависимости от цлелей исследователей. Записи не только аудиоматериала, но и видеоматериала и их дальнейшее дигитализирование позволили бы исследование в будущем и невербальных компонентов коммуникации.

Область изучаемых явлений – точка пересечения теоретической и прикладной лингвистики и множества интердисциплин – психо-, этно-, социо-, контрастивная, корпусная лингвистика.

Выбор **методики** сбора материала определяется целями и объектом исследования. Записи автентичного аудиоматериала (а в перспективе и видеоматериала) собираются с помощью метода включенного наблюдения с помощью скрытого микрофона.

"В лингвистической и методической литературе **билингвизм** определяется как особый случай языкового контакта, при котором проявляется способность человека употреблять для общения две языковые системы. Эта способность осуществляется при помощи речевых механизмов, позволяющих индивидууму переключаться с одной языковой системы на другую [Бабов 1974: 7]. Нужно выявить факторы, влияющие на выбор одного из языков – социальная сфера общения, коммуникативная ситуация, отношения между участниками, моментное эмоциональное состояние коммуниканта и др. Нас интересует, в частности, и явление, обозначенное в литературе как "переключение кода". Оно может принимать разные формы: выбор одного из языков, чередования языков в длинных высказываниях, начало одного высказывания на одном, а заканчивается на другом языке, единичные "вкрапления" из другого языка и др. Это явление рассматривается наряду с заимствованными словами, языковым взаимодействием, языковой конвергенцией – это лишь некоторые из проявлений языковых контактов [Савич 1996: 13].

В наших записях есть лица как с одинаковым совершенным владением двумя языками – одним из видов билингвизма, так и лица с неодинаковым владением двумя языками, применяемыми в разных условиях – двуязычие как диглоссия [разграничения О.С.Ахмановой 1975: 98]. Собранный корпус должен охватывать как исследуемые лица с естественным билингвизмом (результат пребывания и общения в иноязычной среде), так и лица с искусственным билингвизмом (или т.наз. "академическим" билингвизмом: задача обучения иностранному языку – "создание" билингвизма у учащихся) [Бабов 1974: 8]. Среди лиц с родным языком болгарским объектом наблюдения является репродуктивный и продуктивной билингвизм (когда изучающие русский язык как иностранный умеют вопринимать и воспроизводить тексты в устной и письменной форме, но и свободно выражать свои мысли на втором языке), а исключаются билингвы с первой, наиболее ранней стадией билингвизма – рецептивным, который предполагает только восприятие и понимание иноязычного звукового и графического текста [см. Верещагин 1969].

Транскрибированные аудиозаписи пока незначительны по объему, но уже предоставляют интересный материал для наблюдения.

"Переключение кода" обусловлено тремя группами факторов: языковыми, социальными и индивидуальными [см. Петрова, Цветкова 2002]. Оторванность русскоязычных, проживающих постоянно в Болгарии, от живой родной языковой среды, речевых навыков и традиций создают фактор смешанного типа.

Среди языковых факторов можно отмеитиь отсутствие эквивалентности между лексическими единицами: Например: *Иду посмотрю #манджа#* (по правилам транскрипци вводим особый знак, указывающий границы "переключения кода"). При семантической эквивалентности между болг. "манджа" и рус. "еда, кушанье", есть разница в стилистической окраске и функционировании лексем в двух языках. Так что переключение кода в данном случае - маркер стилистической принадлежности болгарской лексемы к "домашней", бытовой сфере (слово является турцизмом в болгарском языке).

Социальные факторы, влияющие на поведение билингвов, определяются сферой общения, отношениями между собеседниками. Как подтверждают наши записи, официальная сфера общения (на работе, в разговоре с незнакомыми людьми, в сфере общественных услуг и под.) нормативность и информативность подавляют экспрессивность. И наоборот, в неофициальной обстановке (дома, в дружеской компании и др.) "нормативизирующий фильтр" снимается, контроль за речью ослабляется, и усиливаются проявления интерференции и переключения кода.

Индивидуальные факторы тесно связаны с социальными, с психическим состоянием говорящего. Один из фактов, которые привлекают наше внимание – это использование служебных экспрессивных слов (междометий, частиц, модальных, паразитных, вводных слов) из родного языка при использовании "другого", иностранного языка, особенно когда говорящий не постиг свободы и легкости общения говорения по сравнению с родным языком. Объяснения можно искать в разных направлениях. Вероятно, контроль за этим типом служебных "сопровождающих" слов, эмоционально нагруженных, выражающих субъективную модальность ослабевает в неформальной сфере общения, в спонтанной и неподготовленной речи, они являются "клапаном" освобождения напряжения и доскомфорта, порожденных ощущением недостаточного владения языком и адекватности коммуникации. Возможно, это иллюстрация психологических исследований о том, что на начальном этапе изучения иностранного языка внутренняя речь идет на родном языке, хотя внешняя речь поддерживается на изучаемом языке [Барсук 1970: 51]. Или можно искать нейропсихологические и когнитивные основания различных механизмов обработки значений (например, эмоционально-аффективных моделей значения, описанных в Стаменов 1987).

Пример (из речи женщины – носителя русского языка, 47 лет, 22 года проживающей в Болгарии):

#ну# ако/ #наприме'р# / не {Ø:е} измита вана {Ø:та}/

'ну / если/ например/ не вымыта ванна/'

Примечания:

- знак, обозначающий "переключение кода" – переход с болгарского языка на русский и наоборот;

/ // /// - обозначение паузы (короткая, более длинная, длинная);

{Ø:е} – опущена форма глагола "быть" для 3 л. ед. ч. наст. вр., нормативной для болгарского языка "е", в русском языке – нулевая; на первом месте ставится зафиксированная форма, после двоеточия в фигурных скобках дается правильный вариант - это часть адаптированной системы для программ TRASA и TRACTOR.

{Ø:та} – опущен определительный артикль "–та" для сущ. ж.р.;

В данном примере в речи на болгарском языке "вкрапливаются" междометие и вводное слово из русского языка. Наблюдаются также типичные проявления интерференции на уровне морфологии – элементы сопоставительного описания языков.

Например:

#ой / дай-ка# {Ø:да} поседна/

' ой, дай-ка присяду / присесть'

В болгарской реплике используются междометие и форма повелительного наклонения с частицей, опущена частица "да" для образования формы, аналогичной инфинитиву в болгарском языке.

"Переключение кода" может вызвано поиском более точного слова в "другом" языке, или невозможностью выразить адекватно определенную мысль. Собеседник "подхватывает" новый язык, "синхронизируется" с коммуникантом.

Зафиксированы проявления интерференции ("отрицательного" переноса элементов одного языка в другой): регулярные фонетические и суперсегментные различия, межъязыковые паронимы, лексико-грамматическаие черты в темпоральной системе глагола, категории возвратности, категории определности / неопределенности (опущение или добавление в ненужном месте), опущение формы наст. вр. глагола "быть" в болгарском языке, инверсированный порядок слов, когда это не оправдано актуальным членением предложения. В словообразовательном отношении калькируется более частотное использование уменьшительных и ласкательных суффиксов, характерное для русского языка. К возвратным болгарским глаголам "приклеивается" постфикс "–ся" (вместо употребления болгарской возвратной частицы "се", которая пишется раздельно от глагола, в пре- или постпозиции).

Например:

*“е-е /# дразнем**ся**# / разбира се //”* вм. *“дразня се”*

‘е-е / раздражаюсь / конечно//’

*“после #подготвят**ся**# всички други материали/ “* вм. *“се подготвят”*

‘потом подгатавливаются все другие материалы’

К этим материалам можно добавить “отрицательный” материал изучающих русский язык как иностранный в болгарской средней и высшей школе.

Исследование речи русско-болгарских билингвов с помощью современных методов комьютерной обработки интересно и полезно для сопоставительного описания двух языков, для улучшения процессов их преподавания, изучения, для улучшения межъязыковой и межкультурной коммуникации.

Литература

1) Алексова К., К. Петрова. Адаптация на шведската система за транскрибиране на корпуси от разговорна реч. – Майски четения, факултетна конференция, 21-22 май 2002 г. (в печати)
2) Ахманова О. Проблемы двуязычия и многоязычия, М.:Наука, 1972.
3) Бабов К. Проблемы интерференции в процессе обучения русскому языку в болгарской школе. София: Народна просвета, 1974.
4) Барсук Р.Ю. Основы обучения иностранному языку в условиях двуязычия. М.: Высшая школа, 1970.
5) Верещагин Е.М. Психологическая и методическая характеристика двуязычия (билингвизма). М.: МГУ, 1969.
6) Гюли, Пиперидис 2003: Voula Giouli, Stelios Piperidis. Corpora and HLT. Current trends in corpus processing and annotation. http://www.larflast.bas.bg/balric/index/index_eng.htm
7) Петрова К., Я. Цветкова. Превключване на кода в устната реч на руско-български билингви. INSOLISO 25-27.10.2002 София, Международна конференция “Съвременни форми на съществуване на българския език”,
8) посветена на 90-та годишнина от рождението на проф. д-р Стойко Стойков (в печати).
9) Савич 1996: Savić J. Code-switching: Theoretical and Methodological Issues. Beograd, 1996.
10) Синклер 2000: Sinclair J., Corpus Processing. 2000 – on-line publications of The Tuscan Word Centre http://www.twc.it/ , To appear in Sterkenberg P (ed) 'A Practical Guide to Lexicography', Benjamins, Final revisions 2002 - текст предоставлен во время курса на факультете славянских филологий Софийского университета, октябрь, 2002.
11) Стаменов М. Семантика на субективното значение, София: Марин Дринов, 1987