Весна ***ПОЛОВИНА***
(*Белград, Югославия*)

МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ СОЗДАНИЯ КОРПУСА РАЗГОВОРНОЙ РЕЧИ

Как нам известно, язык обладает разнообразными манифестациями. Лингвистика подходит к ним посредством разных методов, индуктивных и дедуктивных, путем интуиции, наблюдения и т. п. В историческом смысле, лингвистика, развиваясь через филологию в прошлом, в качестве своего исходного предмета имела письменный язык. Таким образом, она образовала сокровищницу данных и возможностей для обобщения. По-другому дело обстоит с разговорной речью, определяющая характеристика которой – голосовой аспект. А тем самым и изучение этого аспекта было связано с развитием аппаратов для ее записи, сьемки. Этот этап начался только в конце 19-го века (Русло). Разговорная речь была связана с возможностями записи построения корпуса.

Самая общая методологическая проблема с корпусом вообще следующая. Поднимается ли при анализе языка как первое тема и цель исследования, а затем создается корпус, или наооборот, создается общий корпус и на его основании производятся разнообразные исследования разных тематик и целей. Данная проблема немного напоминает классическую проблему: что старше – курица или яйцо? И ответ такой же.

Вторая общая методологическая проблема создания корпуса – во взаимоотношении лингвистической теории и ее отношении к конкретным фактам языка, предоставляемым корпусом. Из истории развития, так называемого корпуса лингвистики в Англии, следует, что отношение к созданию корпуса прошло три этапа: первый до 60-ых годов, затем второй, когда за исходную принимается идея о бесконечном количестве высказываний, которые языковой меха-

низм может произвести, следовательно и бессмысленности корпуса, состоящего из любого конечного количества предложений, и третий этап, свидетелями которого мы являемся.

В этом развитии выявился ряд проблем, которые Лич рассматривает исключительно через роль лингвиста в создании корпуса. Не принимая во внимание технические проблемы или социальные условия, бесспорно оказывающие влияние, как, между прочим, и во всех областях деятельности человека, он свидетельствует о положении корпуса лингвистики в области английского языка, Это, конечно, важно, учитывая распространенность английского языка в мире, и говорит о 17 центрах создания корпуса английского языка, создавших настоящее богатство как материалов, так и опыта в связи с методологическими проблемами.

Считается, что создание каждого корпуса проходит три этапа. Сначала идет запись, в которой максимально поднимается проблема транскрипции и ее инпута. В этой части возрастает сейчас роль компьютера. Уже здесь поднимается вопрос объёма, формата корпуса. Он считает, что сам объём корпуса не имеет такого большого значения, а намного важнее назначение корпуса. Он должен быть репрезентабельный и общий, но и предназначенный для некоторых отдельных функций, например, корпус для языка промышленности и т. п. Кроме того, существует проблема взаимоотношения между письменным и речевым корпусами, затем проблема быстрого развития технологии и, в то же время, медленного развития общественных учреждений, затем относительно быстро развивается хардвер, а софтверные программы развиваются намного медленнее. Указывается на необходимость аннотированных корпусов, важность разницы между омонимами и омографами. Считается, что объем корпуса зависит, т. е. его следует рассматривать через разделение работы между аналитиком, корпусом и софтвером. Мегакорпусы могут быть пригодными для тестирования и изучения разнообразных моделей языка. Статистические данные также могут показать значение одной грамматики, т. е. языковой модели, особенно когда речь идет о пробабилистических моделях языка.

Аннотирование корпуса происходит до его использования, и здесь происходит взаимодействие между человеком и машиной, прав-

да, более точно было бы, если бы он сказал между – лингвистом и электроником.

При аннотировании грамматические и лексические категории можно отметить. Если машина допускает ошибку, опять же таки лингвист и инженер предлагают оригинальную группу категорий или программу, либо и то и другое.

Затем, Лич рассматраивает цель и методы, считая что состояние на сегодняшний день должно закончиться будущим, перспективами развития корпуса лингвистики. Напоминается, что пока существует корпус на 16 европейских языках. Он выделяет французский Тресор де ла лангуе с огромным историческим корпусом, а взаимоотношение между письменным языком с 365 миллионами слов и разговорной речью с 16 миллионами слов.

Проблема траскрипции разговорной речи неразрешена. Как подойти к группе стандардов по энкодированию и транскрипции разговорной речи. Иногда специалисты по фонетике должны это решать и для нужд лексикологов. Очевидно существует недостаток специалистов по фонетике, которые бы помогли при создании корпуса разговорной речи (такие корпуса, несмотря на то, что разговорные почти неиспользуемы для фонетического анализа, который бы должен быть очень существенным.

Лич также считает, что аннотации не могут остановиться на синтаксическом анализе корпуса. Следующая задача – семантический и дискурсный анализы в развитии аннотаций корпуса. Повальное вмешательство человека в создание аннотированных корпусов заключается в первоначальном периоде и в диагностике ошибок. Многие софтверные средства существуют только в форме прототипа, без соотвествующей документации или явного доступа.

Аннотированные корпуса представляют возможность анализа, который не ограничивается на исследовании первоначальных аннотаторов. Необходима более подробная документация о лингвистических схемах, включенных в аннотации, схемы для аннотации.

Указывает на фонетику, морфологию, просодию, мало используемые в корпусах, и семантику и прагматику. Но оригинальный, сырой, необработанный корпус может быть интересным для тех, кто будет считать, что необходимо раскрыть текст в его невинной чистоте.

Одной из основных методологических проблем при создании корпуса является немонолитность лингвистики. Авторы корпуса, аналитики корпуса из числа английских специалистов, например, говорят лишь об корпусе английского языка, не упоминая вообще остальные, сформировавшиеся за пределами их кругозора. А ведь на востоке, в славянских странах, почти параллельно с английскими, создаются в Болгарии, Югославии, Польше и других странах свои корпусы на основании весьма широких исследований. В Югославии – это **Новисадский корпус**, предназначенный, в первую очередь, психолингвистическим исследованиям дискурс анализа, который включает и записи и аннотации самых разнообразных текстов. Этот корпус создается под научным менторством проф. Свенки Савич с 1977 года до наших дней, значит целых 20 лет. В Белграде, в течение восьмидесятых годов, создался статистически релевантный корпус сербского разговорного языка в среде образованных людей. Это результат необходимости исследования текстуальных тем на сравнительно-мультилитературной основе с целью выявления некоторых общих характеристик кохезии текстов на английском, французском и русском языках.

В лингвистике особо выделяются такие исследования разговорного языка в России. Они получают поддержку разных государственных институтов, выпущено немало книг на эту тему (напр. Земская, **Русский разговорный язык, Тексты**).

Весьма интересен и корпус Болгарского разговорного языка, создававшийся в Велико-Тырновском Университете с 1978/1979 гг. Из этого корпуса были уже в 1985 году собраны записи разговоров в объеме 330 часов. В сборниках **Проблемы болгарской разговорной речи** с 1990 по 1996 гг. весьма широко разработаны проблемы, связанные с этой темой, а заметна и тенденция к обобщению проблематики разговорной речи.

Это, конечно, не всё, что можно было бы упомянуть в рамках нашей темы в связи и с другими языками, что было бы весьма полезно и для других, хотя, к сожалению, этим опытом мало кто пользуется.

Создание корпусов разговорной речи в Сербии, Югославии, имеет двадцатилетнюю традицию. Оно совпадает с заинтересованностью небольшой группы лингвистов текстуальной лингвистикой

и анализом дискурса, и таким образом, записанный и транскрибированный материал первоначально был использован преимущественно для таких анализов. Таким образом, индивидуально создаваемые материалы были в зависимости от темы, метода и концепции обработки темы в разговорной речи. Например, работа, направленная на изучение лексической кохезии в разговорной речи в нескольких языках, обуславливала предварительный анализ корпуса на этих языках, из которых ни один не обеспечивал удовлетворяющие элементы для обработки темы. Необходимо было решать некоторые проблемы аннотации и делимитации записанных разговоров.

Такие разнородные методологические проблемы при создании соответствующего корпуса языка поднимались и перед исследователями в течение последних двадцати лет в Югославии. Многие из них решались по мере необходимости отдельных исследователей. Из такого нашего опыта мы можем указать на следующие методологические вопросы создания корпуса разговорной речи.

Первая проблема относится к типу разговора, который необходимо записать для получения данных, существенных для выбранной темы. В рамках данного вопроса, относительно мало внимания уделялось типу говорящего, т. е. сколько общих черт разговаривающие должны иметь в определенном корпусе. Не всегда достаточно принять критерий возраста, образования и синадрадного городского стиля, как принято в уже записанных разговорных корпусах. Это не обязательно должны быть единственные критерии при анализе определенной темы.

Также известна проблема скрытой записи. Записывать без знания говорящего, конечно, лучше, но с его последующим согласием и совместным прослушиванием записи.

Часто подразумевается, что с целью дальнейшего использования материала необходимо имена говорящих заменить другими, выдуманными именами. В транскрипции материалов присутствуют разные решения. В некоторых корпусах указываются только начальные буквы имени, что не совсем удобно, так как в разговоре говорящие обращаются друг к другу не только полным именем, но и из него строятся гипокористики, аугментативы и т. п.

Иногда вместе с аудиолентой (а видеозаписи намного реже в наших корпусах) необходимо обозначить и некоторые действия, соп-

ровождающие жесты, мимику лица, смех, и т. п. В существующих корпусах и эта проблема не решена лучшим или консистентным способом. Например, если в скобках будет написано смех, из этого не следует, кто смеется, все или отдельные лица. Вместо общего имени существительного, более полезно использовать глагол *смеются, смеется*.

И другие, принятые сокращения, не всегда достаточно информативные. Например, *нрз.* в значении "непонятно" может относиться к непонятности из-за того, что несколько человек говорят вместе, но и по другим причинам, например, вследсвие определенного ситуационного прекращения разговора (угощение, шум вследствие передвижения предметов в ситуации и т. п.).

Во всех корпусах, которые мы рассмотрели до настоящего времени, несоответственно решены дискурсные элементы. Это – проблема сверхфразовых целостей, абзацев и их графической презентации. Этому бы соответствовал лингвистический подход, по нашему мнению, упрощенный, согласно которому диалог является речевой цепью, постоянным континуумом, который в письменной форме невозможно сегментировать. Такая графическая презентация корпуса не должна мешать читателю, который занимается грамматическо-морфологическими элементами в рамках предложения или реплики, а именно это и есть то, что до сегодняшнего дня больше всего изучалось в разговорной речи. Такая схематизация не может удовлетворить при анализе других уровней разговора, тематическая структура которых выходит за рамку предложения. Но разговорная речь содержит и целости более широкие, чем реплика или парареплика. Единственное возможное решение – это разговорные абзацы, аналогичные письменным формам абзацев.

В связи с этим мы считаем, что в транскрипции разговорной речи необходимо использовать как можно больше знаков препинания, какие постоянно встречаются в текстах прозы, а именно для всех явлений в разговоре, учитывая навыки читателей, что облегчает понимание. Например, мы считаем что две или три точки для обозначения паузы уже легче для понимания, чем, скажем, одна, две, три перпендикулярные черты. Или зачем избегать использования вопросительного знака, принятого в транскрипции, чтобы обозначить предложение с вопросительной грамматической или только вопро-

сительной интонацией. Таким образом, предотвратилось бы образование множества новых знаков для пауз и интонации, задержалось бы то, что представляют орфографические правила, которые также основаны на логике. Новые знаки в транскрипции потом могут добавляться для обозначения новых явлений, которые иначе не обозначены в препинании текста, а необходимы для транскрипции разговорной речи.

Паузы представляют отдельную проблему для обозначения в корпусе. В разговоре существуют таковые, которые не представляют один и тот же вид, как те, которые стандартная орфография обозначает с помощью запятой, двоеточия и т. п. Такие приостановления имеют функцию раздумывания, самокоррекции, поиска соответствующего выражения, воспоминания и т. д. и появляются там, где их логическая интерпункция, в основном, не могла бы выразить соответствующим образом. Такие короткие приостновления могут обнозначиться с помощью одной наклонной черты, а более длинные – двумя наклонными чертами. Продолжительность одной наклонной черты немного длиннее, чем это обозначает запятая. Правда, в некоторых корпусах она используется вместо запятой. Две наклонные черты обозначают более длительную паузу. Конечно, крайне точное обозначение было бы обозначение продолжительности паузы указанием количества секунд. Но это важно для точного фонетического анализа, а не для всех видов исследования, так что такое обозначение затрудняло бы других исследователей, замедляло бы чтение корпуса.

Также, введение других знаков, треугольников, четырехугольников, звездочек и т. п. может затруднять непрерывность восприятия текста. Часто идет речь не о знаках международной фонетической транскрипции, а о традиции в рамках одного корпуса, что также затрудняет международное использование таких текстов. Задача символизации и есть упрощение схематического представления определенного сложного феномена, а не его усложнение.

В корпус разговорной речи важно ввести обозначение смены говорящих, способ передачи слова, вторжение в речь другого или записать совпадение одинаковых или почти одинаковых высказываний говорящих. Такие явления имеют по-разному решенное графическое оформление в разных корпусах. Так, например, в русском

образце: /*Угу*/ – короткое фатическое высказывание говорящего вписывается в реплику говорящего в этот момент, в английском корпусе обозначается звездочками одновременность частей высказываний двух говорящих. Существует и третья возможность, которую мы использовали, а именно нелинеарное обозначение, посредством вертикального графического решения, либо и чертой увязывание совпадающих высказывании. Например:

А: – *Ты сегодя пойдешь туда?*

Б: – *Ух, ух!*

А: – *и спросишь. Она тебе все скажет.*

Или:

А: – *Некоторые говорят это смешно.*

Б: – *Кроме…*

А: – *Кроме того она хотела*

Иногда случается, что двое или несколько говорящих в одно и то же время дают абсолютно одинаковое высказывание. Например:

А: – *Петр уже пришел?*

А и Б: – *Пришел.*

Можно проставить обе начальные буквы обоих имен двух говорящих.

Также поднимается вопрос включенных частей в разговор, которые определенным образом представляют дигрессию, непредусмотренное прерывание и т. п. Это бы тоже могло быть решено обратными угловыми скобками. Например, перебивание:

Н: – *Тогда я пришла домой, но знаешь … затем комментарии, разное, разное. Все это, говорю тебе, вообще ненормально!*

]В: – *Так я, прихожу к Мите на работу, и совсем другой разговор, чем у меня был.*

Н: – *Встретили меня несколькими анекдотами…*

Еще одна очень важная сторона корпуса решена не совсем удачно. Это описание ситуации и существенных ситуационных факторов. Такое описание должно бы, уже при грубой транскрипции, быть более обширным и с учетом места, на котором велся разговор, времени и т. п. Проблемы, в связи с этим недостатком, легко замечаемы, даже если материал транскрибирует человек, у которого от-

сутствуют такие информации и когда такой материал получает новый потребитель для обработки своей темы. Например: *Я повернусь спиной к Пекину* не является метафорическим политическим выражением, а разговор происходит в офисе, выходящем на ресторан с китайской кухней с названием *Пекин*. Такие примеры, но в более широком смысле, если речь идет о социолингвистических и психолингвистических анализах корпуса, указывают на необходимость внесения намного больше данных о ситауционных факторах.

Кроме таких информаций, при создании корпуса методологически важно принять решение и касательно вопроса отношения к степеням лингвистического анализа разных уровней разговорной речи – фонетического, морфологического, синтаксического, семантического и текстуального. Проблемы кодирования корпуса в отношении грамматических видов слов и грамматических категорий для флективного языка, как сербский, являются, конечно, комплексными проблемами, над которыми до сегодняшнего дня больше всего работали. Омонимия в письменной форме языка также чаще присутствует, чем в разговорной речи, учитывая роль лексического ударения в сербском языке. Известны и специфичные разговорные феномены дискурсных маркеров, таких как дискурсное употребление деиктиков *овај, тај,* … частиц *бре, као* и т. п., фатических глаголов *знаш* и т.п., статус которых в грамматическом смысле трудно определить. Синтаксический анализ также выдвигает своеобразные проблемы перед аналитиками разговорной речи и определенные иные отношения к взаимосвязанности между языковыми уровнями (синтаксиса и просодии, например). Несмотря на то, что степень анализа отдельного языкового уровня можно определить только в зависимости от рассматриваемой темы о корпусе разговорной речи, это не значит, что другие уровни не могут быть существенными в определенный момент.

Можно заметить, что, за исключением отдельных корпусов в мире, мало внимания уделялось фонетическому и фонологическому аспекту разговорной речи, что, несмотря на то, чем вызвано, считаем, что должна быть одной из первичных задач обработки корпуса данного типа. Также семантический и дискурсный анализ, включая и ситуационные факторы, становятся все более значительными в наших исследованиях.

В общем рассматривая, основные методологические проблемы создания корпуса разговорной речи имеют свою практическую сторону: самые экономичные и самые простые нотации в транскрипции и кодировании корпуса для использования, и какие пояснения необходимы в аннотированных корпусах. Принятие решения по данной проблеме обусловлено и дальнейшим усовершенствованием теоретического подхода к разговорной речи.

В институциональном смысле определенные существующие опыты указывают на то, что существует возможность более широкой организации и согласования практических и теоретических решений, по примеру диалектологов, которые организовали разработку европейского диалектологического атласа.

Одной из наиболее важных задач при решении проблем методологии создания корпуса является преодоление различий в лингвистике, в ее максимальной интеграции. Первым шагом в эту сторону мог бы быть новый методологический прием, который бы давал возможность до записи корпусов изучить все, что до сих пор осуществлено в этой области в ракурсе пространства и времени. Затем, необходимо применять все лучшие из решений, которые бы являлись основой для исследователей языка в собственных работах, в рамках того, что им предоставляет современная техника. Таким образом применялся бы всеобщий прием, и на основании сформировавшихся международных стандартов разрабатывался бы дальше корпус, имеющий огромное значение в лингвистическом анализе.

ЛИТЕРАТУРА

Aarts, J and Meijis, W. (eds.), 1990, *Theory and Practice in Corpus*, Amsterdam, Rodopi.

Aktenberg, B. 1991, *English Corpus Linguistics, Studies in Honour of Jan Svartvik*, London–New York, Longman.

Armstrong, S. 1994, *Using large corpora*, Cambridge, Mass. MIT Press.

International Journal of Corpus Linguistics, 1996, Amsterdam, Philadelphia, no. 1.

Йосифова, Рашка. 1985, *Упътване за събиране на материал от книжовно-разговорната реч*, Велико Търново, Филологически факултет.

Johanson, S. and Stenarom, A. 1991, *English Computer Corpora*, Berlin, Mouton de Gruyter.

Leech, G. 1991, *The State of art in corpus linguistics*, in: Aktenberg, B. English Corpus Linguistics.

Polovina, V. 1984, *Srpskohrvatski razgovorni jezik (Tekstovi)*, Dodatak doktorskoj disertaciji: Leksicko-semanticka kohezija u razgovornom jeziku, 1985, Beograd, Filoloski fakultet.

Polovina, V. 1987, "Some problems in the segmentation of spoken conversational language", *Proceedings of the XIV International Congress of Linguists*, Berlin, 1987, III, 2198–2220.

Проблеми на българската разговорна реч, Велико Търново, 1991, 1994, кн. 2; 1995, кн. 3; 1998, кн. 4.

Русин Русинов, 1991, *Проблеми на българската разговорна реч*, Велико Търново, Приветствие при откриване на научната сесия, 5–9.

Savic, S. i V. Polovina. 1987, *Srpskohrvatski razgovorni jezik*, Novi Sad, Filozofski fakultet.

Savic, S. i V. Mitro. *Diskurs telefonskih razgovora*, Novi Sad, Futura publikacije.

Vitas, D. i C. Krstev. 1992, "Interaction between Dictionary and Text in Serbo-Croatian, *Papers in Computational Lexicography*, Complex'92, Linguistic Institute, Hungarian Academy of Science, Budapest, pp. 333–342.