ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 44, КН. 1, СБ. А, 2006 – ФИЛОЛОГИЯ
PLOVDIV UNIVERSITY „PAISII HILENDARSKI“ – BULGARIA
SCIENTIFIC PAPERS, VOL. 44, BOOK 1, PART A, 2006 – PHYLOLOGY

БЪЛГАРСКАТА РАЗГОВОРНА РЕЧ В ИНТЕРНЕТ

Йовка Тишева
Софийски университет „Св. Климент Охридски“

В последните години в резултат на все по-масовото навлизане на интернет технологиите в живота на различни възрастови групи съвременният български език получи трайно представяне и във виртуалното текстово пространство. Това важи както за писмените текстове от различни стилове, сфери на употреба и жанрове, така и за спонтанната неподготвена реч. Устната форма на съвременния български език получи специфична реализация в интернет чрез форуми, чат, различни групи по интереси, виртуални клубове и под. Наред с това вече са направени и успешни стъпки за публикуването на транскрипции и аудиофайлове на спонтанна устна реч в неофициална среда. Тези нови реалности, свързани с функционирането на българския език, променят не само представата за написан и произнесен текст. При анализите на езиковите явления наред с позоваването на класическите литературни текстове все по-често езиковедите търсят данни и от сферите на устната реч. Необходимостта от задълбочено представяне на българската разговорна реч е свързана на първо място с разширяване и стандартизиране на данните за нея.

В тази работа[1] ще бъдат разгледани някои спорни въпроси, свързани представянето (публикуването) на данни за българската разговорна реч в интернет – структуриране на данните, избор на формàт за визуално представяне, подбор и йерархия на елементите, които да

[1] Анализите, представени в статията, са част от работата по научноизследователския проект „Поддържане и обновяване на базата данни за съвременния български език“, етап II (договор № 95 от 14. 04. 2006 г. на СУ).

бъдат анотирани. Създаването на езикови корпуси се прави и с идеята те да подлежат на компютърна обработка. Оказва се, че създадените компютърни средства за работа върху стандартни корпуси (от писмени текстове) не са приложими при транскрипции на записи „жива" реч. По-приемливо е, вместо да се разработват специални средства за обработка на данни, представящи българската разговорна реч, в транскрипциите – в една или друга степен – да се прилагат начини на записване, близки до кодифицирания правопис, до записването в стандартните корпуси.

При представяне на особености на разговорната реч трябва да се вземат предвид не само ярките ѝ отлики от стандартния език на равнището на фонетиката и граматиката. Тъй като устната реч се реализира предимно в диалогична форма, тя притежава специфични особености и на равнището на диалога и дискурса, на прагматиката и комуникативните стратегии. В корпусите на българската разговорна реч тези отличителни характеристики задължително трябва да бъдат запазени.

ЗАЩО БЪЛГАРСКАТА РАЗГОВОРНА РЕЧ ДА СЕ ПРЕДСТАВЯ В ИНТЕРНЕТ

Преди лингвистичните анализи на тази форма на езика ни обаче все пак трябва да се потърси мотивацията за създаването и поддържането на подобни текстови ресурси. Макар въпросът за това, защо българската разговорна реч да се представя (и) в интернет, да звучи вече почти риторично, тук ще бъдат посочени само някои от възможните приложения на публикуваните в интернет данни за устната реч. На първо място, те представляват задължителният паралел на корпусите от писмени (стандартни) текстове. Така ще се допълва и обогатява информацията за съвременния български език чрез представяне на най-динамичната му форма. В групата на потенциалните потребители на тези корпуси най-голям, разбира се, ще е броят на тези, които се нуждаят от надеждни данни за лингвистични изследвания както в областта на съвременния български език, така и за съпоставителни изследвания в по-широк – в балкански, славянски, типологичен – план. Подобни корпуси (особено ако в тях са включени освен транскрипции и аудио- и видеофайлове) могат да подкрепят обучението по български език като чужд с представяне на реалното „битие" на изучавания език, различните езикови стратегии, някои от невербалните средства за комуникация. Метаинформацията, която се

включва към транскрипциите – социолингвистични данни за говорещите, за условията, при които протича комуникацията и под. – правят корпусите на разговорната реч приложими в редица области на хуманитаристиката и в обществените науки – социология, психология, теория на комуникацията. Данните биха могли да се използват и за целите на езиковото моделиране, синтез на реч, машинен превод и др.

При създаването на подобен тип корпуси идеята за представителност на данните, които ще бъдат включени, е водеща. От друга страна, макар че „нито един корпус не е може да се разглежда като „представителен модел" на целия език, подобни ресурси се създават с идеята да представят максимален брой езикови особености" (Стъбс 2001: 305). Постигането на баланс между стремежа за натрупване на максимално количество транскрипции (часове записи) и търсенето на разнообразие както в речевите ситуации, така и в чисто лингвистичен аспект би довело до създаването наистина на надеждна и представителна база данни. Началният етап от създаването на корпус на българската разговорна реч е свързан, разбира се, със събиране на текстове, без да са поставяни определени условия за мястото, участниците, темите на разговорите. Това, което обединява събраните данни, е, че те представят устната форма на съвременния български език в спонтанното всекидневно общуване предимно в неофициална обстановка.

В настоящия момент вече са натрупани и достатъчно теоретични знания за граматиката на устната форма на българския език, затова от съществено значение се оказва въпросът за това, кои черти на разговорната реч задължително да бъдат илюстрирани с текстове (и дори синхронизирани със аудиозаписите). Следващата стъпка към обогатяване на корпуса е анотирането му. То отново налага предварителен теоретичен анализ, чрез който да се систематизират езиковите особености, които да бъдат специално отбелязани, така че да са достъпни при автоматично търсене и обработка. Анотирането на данните, маркирането на определени морфосинтактични значения на думите (тагирането) обикновено е мотивирано от конкретните изследователски цели. Корпусът, разработван от екип от Факултета по славянски филологии на СУ „Св. Климент Охридски", не е свързан строго специализирани цели (изследване само на определено езиково равнище или на определен езиков феномен). Броят и видът на

езиковите особености, които да се анотират, както и последователността, в която това да се прави, се подчиняват на идеята, че устната реч се реализира в диалогична форма и затова обработката може да започне от равнището на прагматиката, а не да следва структурата на езиковата система и да започва от фонетичното равнище.

ДАННИ ЗА БЪЛГАРСКАТА РАЗГОВОРНА РЕЧ, ДОСТЪПНИ В ИНТЕРНЕТ

Публикуването в интернет на текстове, представящи българската разговорна реч, вече има своята сравнително дълга история. Преди анализите на някои от специфичните характеристики на устната форма на езика, които трябва да се отразят в транскрипциите, ще бъде направен кратък преглед на съществуващите ресурси.

– www.hf.uio.no/east/bulg/mat

На този адрес се намира първата публикация на данни за българската разговорна реч в интернет. Там са представени данни за устното общуване в софийски семейства – архивът на Кр. Алексова, събиран в периода 1989-90г.; част от транскрипциите на Цв. Николова, записите за които са от периода 1975-77; транскрипции на заседания на Великото народно събрание от октомври 1990 (в този период все още няма практика самата институция да публикува материали за работата си или подобни записи; днес такъв тип информация е достъпна на сайта на Народното събрание) ; IRC (Internet Relay Protocol) – материали, предоставени от М. Джонова, 2001, които дават информация за развитието на българската разговорна реч в условията на компютърно опосредствана комуникация.

– www.ling.gu.se/~leifg/sofia

Данните, публикувани на този адрес, представят резултатите от международния проект „Multimedia and Research on Multimodal Social Interaction“, разработен от изследователи от Факултета по славянски филологии на СУ и от Университета в Гьотеборг, Швеция. Там са събрани освен транскрипции на аудиофайлове и транскрипции на видеофайлове. Целта е да се изследват по-голям брой елементи на устната комуникация, като чрез видеоинформацията се отчитат и невербалните елементи. За съжаление, обработката и екстракцията на данни от този архив става с помощта на програмите TRASA и TRACTOR, които не са отворени за свободно ползване.

– www.bgspeech.net

Хронологически последният опит за публикуване на данни за българската разговорна реч съдържа материали, събирани и транскрибирани от различни екипи от Факултета по славянски филологии на СУ – студенти от бакалавърските и магистърските програми, докторанти, преподаватели. Информацията е структурирана според това, с каква речева ситуация се свързва (по фактора място). Там са представени и част от публикациите в областта на българската разговорна реч на някои от участниците в екипа.

Прегледът на съществуващите ресурси в интернет за българската разговорна реч показва, че те в голямата си част вече са с исторически характер. Материалите са събирани отдавна и не отразяват най-актуалното състояние на най-динамичната форма на българския език. С изключение на публикацията на Факултета по славянски филологии на СУ записите не се обновяват. От друга страна, наличните ресурси могат да се определят по-скоро като бази данни, а не корпуси, тъй като липсва анотиране на публикуваните текстове. По-точно би било да се представят като текстови архиви, а не като корпуси.

ОСОБЕНОСТИ НА РАЗГОВОРНАТА РЕЧ

На равнището на фонетиката, граматиката и текста в разговорната реч последователно действат две основни тенденции: тенденция към съкращаване (елизии, прекъсвания) и към натрупване (удвояване, многократност). Специфичните особености на устната реч на различните езикови равнища до голяма степен са резултат от тези тенденции. При транскрибиране и на по-късен етап – при анотиране на записите, се регистрират резултатите от действието на едната или другата тенденция, а интерпретацията на съответния факт е по-скоро задача на изследователя, а не на транкрибиращия (анотиращия).

Тъй като проблемът за елизиите и елипсите е дискутиран многократно при анализите на разговорната реч, тук тези явления няма да бъдат обект на коментар. Важно е обаче да се посочи, че те имат устойчив характер и постоянно функционират като маркери на разговорност и спонтанност, означават желанието на говорещия да смени стила или регистъра в хода на комуникацията.

На равнището на граматиката на разговорната реч като устойчиво явление трябва да се посочи унификацията на граматически средства, стремежът да се намали разнообразието или вариативността на маркерите за определено граматическо значение. Така се стига до

намаляване на средствата за изразяване на определено граматическо значение; генерализира се маркерът, който е бил по-широко използван, а тези, свързани с ограничена група думи или с някои изключения, естествено отпадат. Така може да се интерпретира „продуктивността“ на окончанието за 1 л. мн. ч. сегашно време *–ме* и засилената му употреба и при глаголи от I и II спрежение:

(1) *не може да го превръщаме в ученически туризъм само защото трябва да пестиме*[2]

По същия начин се преразпределят граматическите показатели за изразяване на минало свършено време – окончанията с гласна *–о-* се свързват с малка група глаголи, затова те постепенно започват да се изместват от по-честотните с гласна *–а-*. В примера в (2) промените продължават със силната редукция на гласната:

(2) *отидъх на старт и като отидъх на старт се появи тва*

Като проява на тенденция към изравняване на изговора на маркерите за определеност за мъжки и женски род единствено число може да се анализират случаите, които следват модела на пример (3):

(3) *до минус 6 са паднали температурите през нощтъ*

Промени се наблюдават и в групата на местоименията. Те са свързани не само с изразяване на определени граматически значения, например при местоименията подлози *кой, който* и т.н. и местоименията допълнения *кого, когото* и т.н. Може да се говори за конкуренция между местоимения от различни групи – притежателни вместо възвратно притежателни местоимения, както е в (4), въпросителни вместо относителни, както е в пример (5):

(4) *вярвам само на мен си; братовчедка ми, децата ѝ католици си ги направи*

(5) *пише във вестика къде го четох вчера*

Разбира се, употребата на въпросителни вместо относителни местоимения е системно регистрирана в българските диалекти, но при съвременното състояние на устната реч и намаляващото влияние на диалектните маркери явлението би могло да получи и друга оценка.

В групата на служебните думи една от най-съществените промени е свързана със системата на комплементизаторите. „Неутралният“

[2] Анализираните примери са от базата дани за българската разговорна реч на Факултета по славянски филологии на СУ, достъпна на адрес www.bgspeech.net.

комплементизатор *дето* може да замени *че*, както е в (6), а освен това и системно да се употребява вместо относителните думи:

(6) *той сега го е яд дето тя яде*

(7) *на тая спирка дет слизаме ли*

При изразяване на непряко допълнение доста често предлогът *на* се изпуща. Ако в предложната фраза е включена пълна местоименна форма, липсата на предлога пред нея води до многозначност. В такива случаи като средство за снемане на многозначността на местоименната форма се прилага механизмът на удвояване на допълнението. Кратката местоименна форма показва как трябва да се интерпретира пълната в примери като (8) и (9):

(8) *Ø него пуфи ли му викат*

(9) *Ø мен пък ми е хубаво*

Граматическата природа на механизма на удвояване на допълнението е обект на редица самостоятелни разработки. Тук е важно да отбележим отново, че то е проява на тенденцията към многократност в разговорната реч. По същите причини в устното общуване доста системно се удвоява и подлогът. Процесът, илюстриран в примерите в (10), също е получил своята лингвистична интерпретация във връзка с означаването на елементите на информационната структура на изказването.

(10) *те нали католиците така правят*
тя ми обясняше женатъ// понежи няам дискетки

Тенденцията към повторения, натрупвания на елементи, които носят един и същ тип информация, се реализира и при многократното изразяване на граматически значения. Това се наблюдава при многократното маркиране на определеността в рамките на една именна група. Механизмът може да се реализира чрез членуване и на опората, и на определението, както е в (11), или с употребата на показателно местоимение и членувано име, както е в (12):

(11) *седалката предната ниска/ колкото и напред*
да я дърпам не виждам

(12) *зорке/ у тая тенджера мининката какво е тва*

При транскрибиране на разговорна реч езиковите особености, коментирани тук, би трябвало да намерят съответно означаване (отбелязване в системата за транскрибиране). Това ще улесни значително работата на следващия етап – по тагиране на морфологичните и синтактичните значения на езиковите единици в текстовете.

СТРУКТУРА НА КОРПУСА

При публикуването на текстове, в които се отразява ситуация на реално устно общуване, наред с въпроса за избор на система за транскрибиране се налага да се определят и начините, по които ще се организират както отделните текстове, така и целият корпус. За сравнение стандартните писмени текстове се организират в глави, параграфи, абзаци, редове. Практиката при транскрипциите на български език показва приложението на два модела. При транскрибиране на диалектни текстове се записва само речта на информатора, затова резултатът е текст, доста близък по структура до наративните (монологични) текстове. При транскрибиране на разговорна реч от значение е речевото поведение на всички участници в общуването, затова се запазва диалоговата структура. Такава структура е избрана за текстовете във всички публикувани досега бази данни за българската разговорна реч.

Следващата стъпка по обработка на текстовия масив е линеаризацията на диалога. Така се означава процесът на сегментиране на речта на всеки говорещ на по-малки единици, подредени по определен начин (вж. Едуардс 2001). Речевата продукция се организира в реплики (turns), изказвания (utterances) и интонационни единства (фрази).

Репликата се състои поне от едно изказване, произведено от един говорещ. Проблематичен е въпросът как да се представят няколко последователни реплики на един и същ говорещ, разделени например от паузи, смяна на мястото и под., но без смяна на ролите (без включване на изказвания на нов говорещ). В наложилата се система за записване на данни за българската разговорна реч такива реплики се представят последователно, като цяло, в един абзац, а не всяка нова реплика на същия говорещ като нов ред или абзац.

В практиката са известни два формата за подреждане на репликите на говорещите. Възможно е за всеки говорещ да се въведе отделна колона, в която да се записват репликите му. Така може лесно да се проследят смените на говорещите, относителният дял на репликите на всеки от тях. Основните критики срещу този начин на записване са свързани с факта, че подреждането задава модел на доминиране на един от говорещите, а това противоречи на идеята за сътрудничество в диалога и прогресивно развитие на темата. За данните за българската разговорна реч се е наложил форматът на последователното записване (подобно на текста на драматургично произведение с означените действащи лица). Така се представя нагледно постъпателното развитие на комуникацията (за организацията на дискурса вж. и Schiffrin 2001).

Репликите са изградени от изказвания на един говорещ. Изказването се определя като последователност от думи на един говорещ, произнесени с минимална намеса от страна на друг говорещ. Важно е да се отбележи, че едно изказване може да съдържа повече от едно изречение.

На пръв поглед въпросът за определянето на границите на изказванията изглежда несъществен. Това е така, ако към текстовия файл има и съответната „звукова илюстрация“ с включването на различните гласове в диалога. Сравнително еднозначно може да се определи началото на изказването на даден участник в комуникацията. Проблемите възникват при точното определяне на края на изказването поради непълнота (елипси) или интонационна, синтактична, семантична незавършеност.

Една от основните причини за нееднозначното определяне на границите на изказванията е свързана със случаите на едновременно говорене. То води до застъпване на различни части от репликите на различни говорещи (overlaps). В класическия случай краят на репликата на първия говорещ съвпада с началото на репликата на втория говорещ[3]:

(13) В: *марио ено високо[русоляво момче]*
Б: *[дето пипаше в] контакта (Ф-смях)*

По-сложен случай е илюстриран в (14), където краят на първата реплика на единия от говорещите съвпада с началото на репликата на втория говорещ, а краят на репликата на втория говорещ съвпада с второто включване на първия говорещ:

(14) Н: *ми точно затва /питах съ // дъли шъ [мога дъ ги продавам]*
Р: *[а идеята е]/да си стуиш ф стайъта и хоръта*
дъ идват дъ ги[купуват ли]
Н: *[Нье]*

Записването на застъпванията по представения начин вероятно е най-ефективно при транскрибиране на разговор с двама участници.

[3] Фразите, които се застъпват, са записани в квадратни скоби.

При включването на повече говорещи е трудно да се определи само от текстовия файл кой кога се включва и чии реплики се застъпват. Така в (15) не е ясно дали втората реплика на Б. съвпада по време с началото на репликата на А или с края й. Не може да се определи и с коя реплика се застъпва втората реплика на В.

(15) Б: *//(Ф-ъ-ъ) това го знам (Ф-смях)//*
В: *та такива работи да се разберем/[хората]*
А: *[а не като] американците имат само [една]*
Б: *аз [сега]*
В: *[да]*

При означаването на застъпване на реплики в транскрибирани текстове може да е приложи и друг модел. В (16)а. Примерът е представен по „класическия" начин със скоби. Началото на репликата на втория говорещ съвпада с част от репликата на първия говорещ:

(16) а. Б: *аз толкова трудно казвам не че [направо]*
не мога да се понасям
В:*[да] // а аз съм мека мария и оттам идват моите проблеми*

За такива „вътрешни" застъпвания се прилага и т.н. многолинеен формат. Застъпващите се реплики са изписани една под друга:

(16) б. Б: *аз толкова трудно казвам не че*
[направо] не мога да се понасям
В: *[да] //а аз съм мека мария и оттам идват*

Този формат е доста по-експлицитен, но отново създава неудобства, тъй като след фазата на застъпване репликите на двамата говорещи не са произнесени едновременно, а са изписани паралелно. Много по-ясно представяне може да се постигне с маркиране на времето на застъпване и синхронизиране на текстовете.

За устните текстове е типична високата фреквентност на непълни (незавършени, прекъснати) изказвания или фрази. Често такива изказвания са резултат от желанието на говорещия да промени начина, по който ще структурира репликата си. Явлението се представя чрез термина корекции (speech repairs); в примерите е отбелязано с (ФС) – фалстарт. Промяната може да е свързана с предлагане на ново начало на фразата, както е в (17) и (18):

(17) А: *и между другото аз (ФС) ъ: той има са имен ден*

(18) Р: *така ли/ чакай да ги...(ФС) поне да го...// а не / не не беше така да се сканира*

Ако говорещият коригира своето изказване, като предлага нова фраза непосредствено след коригираната фраза, говорим за модификационни корекции. При съкратените корекции се променя само отделна дума след пауза или запълващ елемент, както е в (19) (за видовете корекции вж. Хиймън и кол. 1997, 1999):

(19) Р: *// амъ това ще ...(ФС)/ това не е: болка за умиране/ най-малкото тукъ могат да се сложат по-неизползваеми дрехи // и приключва проблема//*
П: *//ама е страхотно че://*

Модификациите на фразите и автокорекциите от различен тип се запазват при транскрибиране. Словоредните модели на такива фрази не следват „класическите" структури и това би затруднило използването на такива примери за целите на автоматичното генериране на синтактично дърво например. Коригираните фрази обаче са един от най-ярките маркери за разговорност на текста – те показват, че той е реализиран спонтанно в ситуация на неподготвено общуване. Устната форма позволява в потока на речта говорещият да редактира, уточнява и допълва изказването си.

ИЗВОДИ

Публикуването на корпуси (бази данни) за българската разговорна реч представя най-динамичната форма на езика ни за най-широк кръг потребители, с различна степен на осведоменост и познаване на славянските и балканските езици и на българския език в частност. Затова от съществено значение е корпусът да бъде надежден и представителен. Доброто познаване на особеностите на устната форма на съвременния български език, теоретичната обоснованост на избраната методика за транскрибиране и представяне на данните в корпуса са основата за високата му надеждност.

Избраната дилогова структура за представяне позволява данните да се използват както за изследвания на равнището на синтаксиса, морфологията и лексиката, така и на по-високите равнища – на дискурса и прагматиката на речта.

За да бъде корпусът лесен за използване от потребители с различни (невинаги и не само изследователски научни) цели, форматът и визуализацията на данните трябва да са ясни, удобни и достъпни, както и да се поддават на машинно търсене и обработка с достъпни програми.

ЛИТЕРАТУРА

Едуардс 2001: Edwards, J.A. „The Transcription of Discourse". *The Handbook of Discourse Analysis.* ed. by D. Schiffrin, D. Tannen & H. Hamilton. 54-75. Oxford: Blackwell, 321-347.

Хиймън и кол. 1997: Heeman, P. A. and Allen, J. F. „Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog." *Proceedings of the 35*th *Annual Meeting of the Association for Computational Linguistics*, 254–261.

Хиймън и кол. 1999: Heeman, P. A. and Allen, J. F. „Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialog". *Computational Linguistics,* 25(4). 527-572.

Шифрин 2001: Schiffrin, D. „Discourse Markers: Language, Meaning, and Context." *The Handbook of Discourse Analysis.* ed. by D. Schiffrin, D. Tannen & H. Hamilton. 54-75. Oxford: Blackwell. 54-75.

Стъбс 2001: Stubbs, M. „Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence". *The Handbook of Discourse Analysis.* ed. by D. Schiffrin, D. Tannen & H. Hamilton. 54-75. Oxford: Blackwell. 305-319.