BGSPEECH И ПРЕДСТАВЯНЕТО НА УСТНАТА РЕЧ В БЪЛГАРСКИЯ НАЦИОНАЛЕН КОРПУС[1]

Йовка Тишева, Марина Джонова (София, България)
Хетил Ро Хауге (Осло, Норвегия)

В статията се разглеждат новите електронни ресурси за устната българска реч. Представят се предимствата при използването на такъв тип данни – свободният достъп до тях (както до самите транскрипции, така и до оригиналните аудио- или видеозаписи), наличието на метаданни за говорещите и за комуникативната ситуация, както и възможността за специализирано търсене в корпуса. Съвременните технологии позволяват на лингвистите да верифицират своите изследвания върху богат емпиричен материал. Съвременните корпуси от устна българска реч са създадени според изискванията за представителност и балансираност, т.е. в тях са включени данни от различни сфери и ситуации на общуване, а от социолингвистична гледна точка участниците в устната комуникация се различават по своите характеристики.

Важно е да се уточни понятието „корпус“. Това е сбор (съвкупност, системна организация) от текстове в електронен формат, преминали през различни етапи на обработка (анотиране), което дава възможност за автоматично търсене и по-нататъшна обработка на данните по определени лингвистични и екстралингвистични параметри.

За българския език съществуват корпуси, които са представителни за съвременното състояние на писмената реч, анотирани са и са достъпни за лингвистични цели[2]. Като пример могат да се по-

[1] Част от материалите, представени в тази статия, са свързани с работата по проекта „Изследване на модели и средства в различни речеви ситуации и сфери на общуването в съвременния български език“, финансиран от Фонд „Научни изследвания“, договор № ДТК 02/11 от 16.12.2009 МОМН. Изказваме благодарност на НИС на СУ, с чиято подкрепа, в рамките на проекта „Мултимедиен корпус на българската устна реч – 2 част“, договор № 86/2014, продължи публикуването на данни за българската устна реч.

[2] За подробно описание на съществуващите електрони ресурси за книжовния български език вж. Коева и др. (2012).

сочат Българският национален корпус (http://dcl.bas.bg/bulnc/начало) и Българският национален референтен корпус (www.webclark.org). Българският национален корпус отразява състоянието на българския език предимно в неговата писмена форма от средата на XX в. до наши дни, като устната реч в корпуса е под 1% (вж. Коева, Стоянова 2009: 140). Този корпус дава надеждни сведения за състоянието на българския език в неговата представителна писмена форма, които намират приложение в изследванията на актуалното състояние на езиковата система, за целите на компютърната лингвистика, на теорията и практиката на превода, както и като емпиричен материал в лекционните курсове по съвременен български език. В рамките на БНК в последните години са включени и паралелни корпуси във връзка с разширяването на интересите в областта на компютърната лингвистика към разработване на многоезикови приложения – машинен превод, извличане на информация от многоезикови ресурси и др.

Българският национален референтен корпус (www.webclark.org) се основава изцяло на писмени текстове. Една от целите на проекта, наред с разработването на средства за автоматична обработка на текстове, е създаването на синтактично анотиран корпус от представителни данни за системата на съвременния български книжовен език. Материалите, включени в ресурса, са отново от писмени текстове – граматики на българския език, публицистични и литературни текстове. Поради спецификата на данните и начина, по който те са представени, ресурсът е подходящ при търсене на материали предимно за изследователски цели, включително и при разработването на иновативни методи за обучение и преподаване на български език.

Корпусите от писмена реч съдържат текстове, които са оформени според нормите на съвременния български език. Предимството на корпусите от писмена реч е, че думите в тях съответстват на книжовния правопис, както и че текстовете съществуват в електронен формат преди включването им в корпуса. Това улеснява обработката на текстовите единици и позволява използването на съвременните компютърни технологии при обработката и анотацията на данните.

За да бъде цялостно проучването на съвременния български език, е необходимо изучаването както на неговата писмена форма, така и на диалектите и на устната книжовна реч. Корпусите от устна реч допълват корпусите от писмена реч в няколко аспекта – от една

страна, сферите на употреба на писмената и на устната реч не съвпадат изцяло (напр. битовата разговорна реч, общуването лекар – пациент, преподавател – студент, медийната реч, общуването по телефона и т.н.). Устната реч е по-динамична и в нея се наблюдават редица фонетични, морфологични и синтактични особености, които не се отразяват в писмената книжовна норма – например елизията на гласни и съгласни звукове, употребата на окончание *-ме* за 1 л. мн.ч. при глаголите от I и II спр. (*четеме, говориме*), разширяването на употребата на т.нар. *ах*-аорист (*четах, отидах*), членната морфема *-тъ* за ж.р. под ударение (*радоттъ, песентъ*). Някои от отклоненията от правоговорната норма, които се отбелязват като характерни за устната реч, са свързани с ятовия преглас (*голема, бел; голями, живяли, тяхни*), с депалатализацията на съгласните (напр. *вървъ, учителъ, лакътъ, затварам*), както и с наличието на дейотация (*онеа* вм. *онея*, *таа* вм. *тая*). За устната реч са характерни и замяната на винителните местоимения с именителни (има ли *някой* тук?; *на кой* да дам списъка?), както и редица разговорни конструкции от типа на *тя я* е страх, *на Мария майка ѝ* е лекарка и под.[3] За да бъдат отбелязани посочените особености, за целите на корпусите от устна реч се създава специална система за транскрибиране, която отчита наличието на отклонения от нормативния изговор.

Структурата на корпусите от устна реч също се различава от тази на корпусите от писмена реч, тъй като в транскрипциите се запазва диалогичната форма на речта. В корпусите от устната реч се включват данни и за останалите елементи на комуникативната ситуация – време, място, участници в разговора, като по този начин се дава възможност на изследователите да разглеждат речта не изолирано, а във връзка с конкретната речева ситуация.

По отношение на българската устна реч съществуват редица електронни езикови ресурси, които са свободно достъпни за изследователски цели, но голямата част от тях са само в текстов формат и не са подходящи за електронна обработка (вж. по-подробно у Тишева, Джонова 2011, Тишева 2014). Като пример могат да се посочат езиковите ресурси, публикувани от Хетил Ро Хауге от Университета

[3] За подробно описание на маркерите на устната реч вж. Алексова (2000), Алексова (2005), Тишева (2013а), Тишева (2013б), Джонова, Велкова (2014).

в Осло, Норвегия (folk.uio.no/kjetilrh/bulg). В тях се включват Корпусът на Кр. Алексова от разговори в семейната среда, Корпусът на Цв. Николова от разговорна реч, както и транскрибираните от Ив. Мавродиева записи на дебати в Седмото велико народно събрание. Подобни са характеристиките и на транскрибираната разговорна реч, публикувана в периода 2001–2004 г. в рамките на инициативата BgSpeech (bgspeech.net/bg/resources/conversations.html).

Прегледът на проучванията в областта на устната комуникация показва, че е натрупан богат архив (аудио- и видеозаписи, транскрибирани текстове), който се обогатява, разширява и допълва. Трудността при създаването на корпус от устната реч е свързана с първичната форма на текстовете – това е устна комуникация, така че оригиналните „текстове“ в този тип корпуси всъщност са аудиозаписите. С цел ексцерпирането на данни от устната реч аудиозаписите се транскрибират, но това е само първата стъпка от създаването на корпуса. За да могат да бъдат пълноценно ползвани данните в корпусите, те трябва да бъдат анотирани. По отношение на устната реч компютърните технологии позволяват синхронизирането на транскрибираната реч с оригиналния запис. Така мултимедийните корпуси, които обединяват транскрибирания текст и съответния аудио- или видеофайл, съдържат данни за цялостната комуникативна ситуация и в същото време данните са анотирани и подходящи за автоматична обработка и за разширено търсене. Синхронизирането със звуковия файл позволява представянето в рамките на корпуса и на някои специфични за устната реч особености, като едновременното говорене, фалстартовете, хезитационните паузи, незавършените изказвания. Диалоговата структура на данните в корпуса от своя страна дава възможност на изследователите да изучават не само морфологичните и фонетичните особености на устната реч на ниво изказване, но и структурата на репликата в рамките на диалога – вземането и даването на думата, смяната на темата, изразяването на съгласие или несъгласие. Разбира се, корпусите от устна реч съдържат и данни за паралингвистичните средства, използвани от говорещите (паузи, жестове, мимики, фонетични паралингвистични средства), тъй като те са неделима част от представянето на устното общуване[4].

[4] За подробно описание на корпусите от устна българска реч вж. Тишева, Джонова (2010), Тишева, Джонова (2011).

BgSpeech – електронни ресурси за устната реч

Създаването на корпуси с надеждна информация за актуалното състояние на българската устна реч, които да се обогатяват и допълват с нови данни, е една от основните задачи на екипа на инициативата BgSpeech. Разработваните езикови ресурси не са свързани с конкретни изследователски проекти (проучване на езикови явления, на общуването в определена ситуация, в определена сфера и под.), а представят устната реч в различни сфери и ситуации – от подготвеното официално публично общуване до спонтанната битова комуникация. Фиг. 1 показва какви ресурси са публикувани към момента от екипа на BgSpeech.

Фиг. 1. Електронни ресурси за устната реч, достъпни на bgspeech.net

Обемът на всеки от четирите ресурса е представен в таблица 1:

Таблица 1. Обем на ресурсите за устната реч, достъпни на bgspeech.net

Ресурс	знаци	думи
Транскрибирана разговорна реч 2001–2004	350 771	71 605
Транскрибирана устна реч	438 895	47 792
Мултимедиен корпус на българската устна реч	1 002 019	166 461
Паралелен корпус	339 275	62 080

Тук ще бъдат разгледани накратко българският Мултимедиен корпус от устна реч (bgspeech.net/bg/resources/multimediacorpus.html),

неговата структура, обем, приложение, както и технологиите, свързани със създаването на корпуса и публикуването му в интернет. Събирането на емпиричен материал за корпуса започва през 2009 г. в рамките на Катедрата по български език към Факултета по славянски филологии на СУ „Св. Климент Охридски“ като част от инициативата BgSpeech. Към настоящия момент дейността на инициативата BgSpeech се координира от Лабораторията за изследване на устната комуникация към Катедрата по български език при ФСлФ на СУ. В създаването на най-новите езикови ресурси участват трима членове на тази катедра – проф. Й. Тишева, гл. ас. М. Джонова и гл. ас. Ат. Атанасов. Член на екипа е и проф. Х. Ро Хауге от Университета в Осло, Норвегия. Корпусът е предназначен основно за нуждите на лингвистите – за изследване на актуалните тенденции в българската устна реч и за верификация на хипотези в научните изследвания, както и като елемент в чуждоезиковото обучение. Създаден е с помощта на програмата EXMARaLDA (http://exmaralda.org), която е свободно достъпна и е предназначена именно за целите на компютърната лингвистика – за създаване и обработка на мултимедийни корпуси. Корпусът е регистриран и достъпен през Метанет (www.metanet.eu) от 21.01.2013 г.

Предимството на мултимедийния корпус е, че той дава възможност на потребителя да възприеме непосредствено цялата комуникативна ситуация, тъй като в него се представят синхронизирано транскрипцията и оригиналният аудио- или видеозапис. Така в съответния звуков или видеофайл е налична допълнителна информация, която не е отразена в транскрипцията, поради изискването за четивност на корпуса (напр. интонация, шумни вдишвания и издишвания на говорещите, някои от жестовете/мимиките, които транскрибиращият не е отбелязал). Както посочва Шмит и др., мултимедийният корпус представлява систематизирана колекция от езикови ресурси, включваща данни от повече от една медия (Шмит и др. 2010:1). Подборът на данните, включени в корпуса, често е свързан с изследването на определен тип общуване – например общуване лекар – пациент, общуване по телефона, медийна реч. Мултимедийният корпус на българската устна реч включва данни от различните регистри на устното общуване. Създаването му не е свързано с изследването на дадено явление в устната реч, а по-скоро цели да представи в

цялост актуалното състояние на българската устна реч, като по този начин да обогати съществуващите ресурси за българския език и да даде възможност за по-голяма пълнота в емпиричния материал, използван при изследването на българския език. Записите са от автентични диалози. Речта е транскрибирана писмено според специално модифицирана система за транскрипция, в която се отчитат особеностите на българската устна реч и се отбелязват някои от невербалните елементи, като мимики и жестове, но предимство се дава на речевото общуване.

Мултимедийният корпус от устна българска реч се състои от цифрови аудио- и видеозаписи от устна реч, синхронизирани със съответните транскрипции на тези записи (текстови файлове). Корпусът включва и два текста, които са синхронизирани със съответния запис с помощта на наскоро разработени на основата на HTML5 технологии. В момента в корпуса са включени 56 транскрипции със средна продължителност на всеки от транскрибираните записи 20 минути. Общият обем на транскрипциите е 1 002 019 знака, или 166 461 думи.

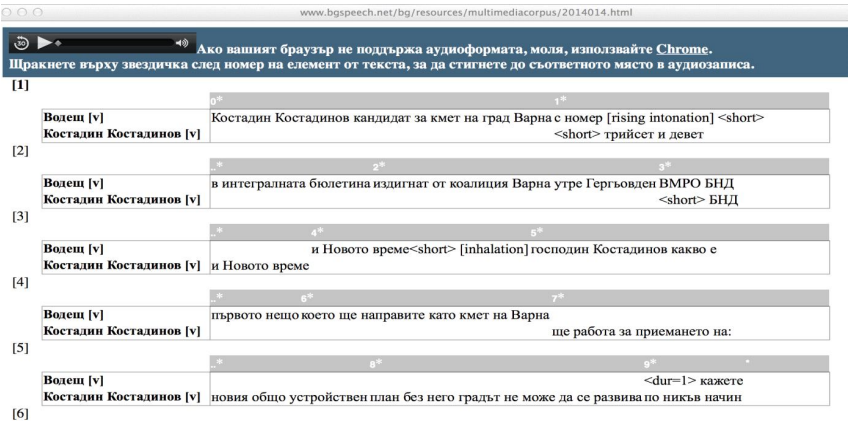
Записите, въз основа на които е изграден Мултимедийният корпус на българската устна реч, са подбрани така, че корпусът да е представителен за устното общуване в различни сфери. Преобладават записите от подготвена публична реч – записи от български електронни медии (радио- и телевизионни предавания), парламентарни дебати, академично общуване и фокус групи. Обемът на данните е показан в таблица 2.

Таблица 2. Обем на данните в Мултимедийния корпус според тематичната област

тематична област/сфера	**знаци**	**думи**
фокус групи	134 562	21 827
медийна реч	718 098	120 442
академична реч	797 730	133 634
парламентарна реч	957 778	158 412

На Фиг. 2 е показано как изглеждат транскрипциите в Мултимедийния корпус. Те се правят с програмата EXMARaLDA (exmaralda.org), която позволява партитурното записване на речта, която от своя страна се синхронизира със съответния отрязък от аудиозаписа.

Фиг. 2. Откъс от транскрипция, включена в Мутимедийния корпус

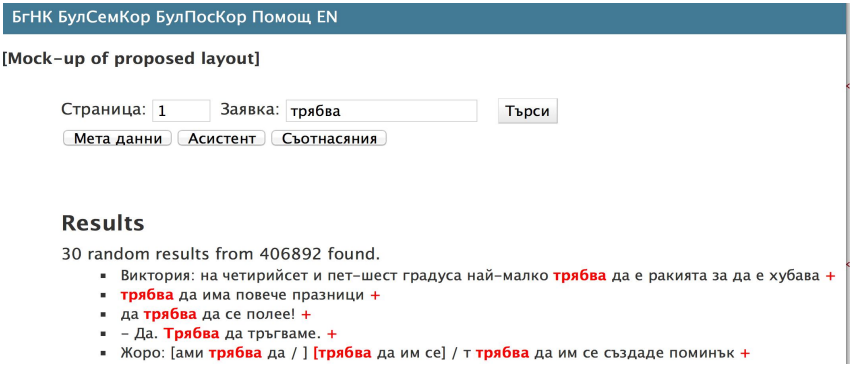


За всеки участник в разговора е налице отделен партитурен ред. Ако в даден интервал от записа е налице едновременно говорене, речта на всеки от говорещите съответства на този интервал от време. Този тип транскрипция позволява на потребителя да *чуе* транскрибираната реч, да избере дали да слуша записа към транскрипцията от самото начало, или да чуе отделни фрази. Изборът на отрязък за прослушване става с натискане на знака * от сивия ред над съответния текст. Избраният начин за записване на речта позволява и отбелязването на невербалните елементи в комуникацията. По-ясно се визуализират типичните за устната реч прекъсвания, вземане на думата или застъпване на реплики и едновременно говорене, тъй като репликите на всеки участник се изписват на отделен ред.

За да бъде пълноценно ползването на корпуса, е необходимо потребителите да могат лесно да извличат търсената от тях информация. За тази цел могат да се използват съществуващите програми за обработка и анотиране на корпуси на български език. Включването на транскрипциите от устна реч в Българския национален корпус ще позволи, от една страна, анотирането на Мултимедийния корпус по части на речта и разширеното търсене в данните, а от друга страна, ще позволи на потребителите на Националния корпус да съпоставят данните от устната реч с тези от писмената реч. За да е възможно това обаче, е необходима нормализацията на текстовете в Мултиме-

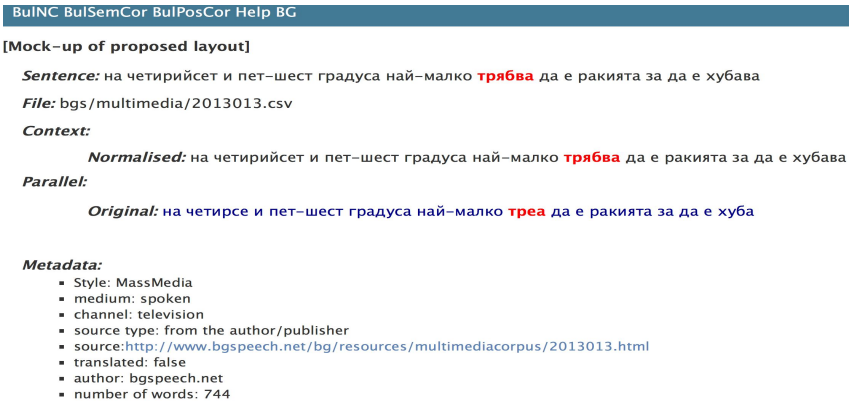
дийния корпус. На фиг. 3 е показано примерно търсене в БНК, при което потребителят е избрал да търси думата *трябва* в данните от устна реч.

Фиг. 3. Резултат от търсене по словоформа в БНК



Във всеки от показаните резултати думата е изписана нормативно – *трябва*. Но при разширяване на резултатите (фиг. 4) се вижда съответствието на думата в транскрипцията, т.е. в резултатите могат да бъдат намерени различни изговорни варианти на думата (напр. *трева, трябва, тряя, треа* и т.н.).

Фиг. 4. Детайлна информация за избраното от списъка с резултати изречение, включваща непосредствения контекст и метаданни за документа източник



Системата за търсене в БНК позволява изследователите да проверяват вариантите на произношение на дадена дума и да правят обобщения за тенденциите в устната реч на базата на бързото ексцерпиране на голям обем от данни.

Може да се обобщи, че развитието на корпусната лингвистика дава нови възможности за ексцерпиране на данни за българския език. Изследователите и преподавателите могат да верифицират своите хипотези на базата на реални примери от езика от различни сфери. В рамките на Мултимедийния корпус са включени данни за актуалното състояние на българската устна реч. Потребителите да имат достъп едновременно до оригиналната звучаща реч и до съответната транскрипция. В транскрипциите се съдържат и данни за паралингвистичните средства, използвани от говорещите (паузи, жестове, мимики, фонетични паралингвистични средства), както и за структурата на диалога – незавършени изказвания, едновременно говорене, прекъсване и вземане на думата. Данните за езиковите явления в устната комуникация разширява представителността на лингвистичните проучвания. Включването на транскрибираната устна реч от Мултимедийния корпус в Българския национален корпус ще даде възможност на потребителите да ексцерпират лесно данни и от устната реч, както и да правят съпоставка между данните за писмената и за устната реч.

БИБЛИОГРАФИЯ

Алексова 2000: Алексова, Кр. Езикът и семейството. Към методиката за проучване на речта в микрообщностите. София: Интервю прес, 2000.

Алексова 2005: Алексова, Кр. Йерархията на социолингвистичните променливи според стратифициращата им сила.// Езиковедски приноси в чест на чл.-кор. проф. д.ф.н. Михаил Виденов. Велико Търново: Унив. изд. „Св. св. Кирил и Методий“, 2005, с. 299–324.

Джонова, Велкова 2014: Джонова, М., Й. Велкова. Фонетични и морфологични особености на устната реч. // Й. Тишева, Кр. Алексова и кол. Как говори съвременният българин – том 1. Граматика и устна реч. София: Фондация „Фокус“, 2013, с. 115–181.

Коева и др. 2010: Koeva, S., Blagoeva, D., Kolkovska, S. Bulgarian National Corpus Project.// Proceedings of LREC-2010. Valletta, ELRA, 2010: 3678–3684.

Коева и др. 2011: Коева, С., Д. Благоева, С. Колковска. Проектът Български национален корпус – резултати и перспективи. // Български език, 58, 2011, № 3, с. 34–53.

Коева и др. 2012: Koeva, S., I. Stoyanova, S. Leseva, T. Dimitrova, R. Dekova, E. Tarpomanova. The Bulgarian National Corpus: Theory and Practice in Corpus Design.// Journal of Language Modelling, 2012, Vol. 0, No. 1: 65–110.

Коева, Стоянова 2009: Коева, Св., Ив. Стоянова. Български национален корпус. // Български език, 2009, № 3, с. 137–145.

Тишева 2013а: Тишева, Й. За разговорните маркери и устната комуникация. // Проблеми на устната комуникация, кн. 9. Велико Търново: Унив. изд. „Св. св. Кирил и Методий“, 2013, с. 73–87.

Тишева 2013б: Тишева, Й. Как говори съвременният българин – том 2. Прагматика и устна реч. София: Фондация „Фокус“, 2013.

Тишева 2014: Тишева, Й. Езикови бази данни, корпуси и електронни ресурси за българската устна реч.// Littera et Lingua, г. 11, 2014, кн. 1–2. http://slav.uni-sofia.bg/naum/lilijournal/2014/11/1-2/ytisheva

Тишева, Джонова 2010: Тишева, Й., М. Джонова. Електронни ресурси за българската разговорна реч (инициативата BgSpeech).// Littera et Lingua, лято 2010, Доклади от научната конференция „Ресурси за електронно обучение“, Факултет по славянски филологии, СУ „Св. Климент Охридски“, http://slav.uni-sofia.bg/naum/node/1735

Тишева, Джонова 2011: Корпус с устна българска реч – структура и специфика. // Български език, 2011, № 3, с. 34–53.

Шмит 2011: Schmidt, T. A TEI-based Approach to Standardising Spoken Language Transcription.// Journal of the Text Encoding Initiative (1), June 2011.

Шмит и др. 2010: Schmidt, T., K. Elenius, P. Trilsbeek. Multimedia Corpora (Media encoding and annotation). Draft submitted to CLARIN WG 5.7. as input to CLARIN deliverable D5.C-3 Interoperability and Standards. at http://www.exmaralda.org/files/CLARIN_Standards.pdf.

Български национален корпус – http://ibl.bas.bg/BGNC_bg.htm
Български национален референтен корпус – http://www.webclark.org
Корпус от устна българска реч – http://bgspeech.net/bg/resources
Корпуси от българска разговорна реч – http://folk.uio.no/kjetilrh/bulg
Метанет – www.meta-net.eu Метанет
Мултимедиен корпус на българската устна реч – http://bgspeech.net/bg/resources/multimediacorpus.html
Сайт за българската устна реч – http://bgspeech.net
EXMARaLDA – http://www.exmaralda.org

BgSpeech and the representation of spoken Bulgarian data in the Bulgarian National corpus

BgSpeech, at bgspeech.net, is a long-term project aiming at collecting and maintaining data on the oral forms of the contemporary Bulgarian language. The project team includes researchers from the Department of Bulgarian Language at Faculty of Slavic Studies of Sofia University and the Department of Literature, Area Studies and European Languages at the University of Oslo. Collection of audio recordings and subsequent transcribing started in 2001, and the first transcripts of data were published on the site in 2004. The present phase of the project is concentrated on making the texts at BgSpeech searchable at the Bulgarian National Corpus (BNC).

BGSPEECH AND THE PRESENTATION OF ORAL COMMUNICATION IN THE BULGARIAN NATIONAL CORPUS

Yovka Tisheva, Marina Dzhonova (Sofia, Bulgaria)
Kjetil Re Hauge (Oslo, Norway)

The article presents the new electronic resources of Bulgarian speech. The accent is on the Multimedia corpus and the advantages of its use for scholarly and research purposes. The possibilities for optimization of the corpus search through the normalization and inclusion of the Multimedia corpus transcriptions in the Bulgarian national corpus are also presented. That would give the users access to advanced search of the corpus.