ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 52, КН. 1, СБ. А, 2014 – ФИЛОЛОГИЯ,
PAISII HILENDARSKI UNIVERSITY OF PLOVDIV – BULGARIA
RESEARCH PAPERS, VOL. 52, BOOK 1, PART A, 2014 – LANGUAGES AND LITERATURE

ЕЛЕКТРОНЕН КОРПУС НА БЪЛГАРСКА ДЕТСКА РЕЧ

Велка Попова
Шуменски университет „Епископ Константин Преславски“

BULGARIAN CHILD LANGUAGE E-CORPUS

Velka Popova
Konstantin Preslavsky University of Shumen

The paper presents bulgarian child language e-corpus developed in the applied linguistics lab at shumen university. Its description focuses on the broad opportunities presented by both the format chosen for the presentation of data (which is actually realized in the terms of the interactive talkbank and childes platform), and the corpus perspective in contemporary linguistics which assures an optimum environment for the establishment of objective language models and restricts the appearance of new myths in contemporary research.

Key words: bulgarian language, child language e-corpus

Увод

Настоящата работа представя електронен корпус на българска детска реч, създаден в Лабораторията по приложна лингвистика към Шуменския университет, като при това е направен опит да се открояят широките възможности както на избрания формат за представяне на данните в него, реализиран в термините на интерактивните платформи TalkBank и CHILDES, така и на корпусната перспектива в съвременната лингвистика, благодарение на която се осигурява оптимална среда за създаване на обективни модели на езика и за ограничаване на появата на нови митове в съвременните научни търсения.

В този ред на мисли напълно естествено възниква въпросът за това дали е оправдано инвестирането на усилия и време в едно толкова трудоемко начинание, при положение че съществуват достатъчно добри алтернативи в традицията. Дали създаването на корпус с детска реч не би могло да се интерпретира като плод на самоцелно слугуване на някаква мода? От какъв род (не)достатъчност „страдат“ познатите вече и добре работещи традиционни модели? В отговор на тези предполо-

жения ще бъде предложен кратък хронологичен екскурс на съществуващите приноси по събирането и организирането на емпирични данни и в този контекст ще бъде потърсено мястото и значението на съвременните електронни формати на корпусната лингвистика.

1. Кратък екскурс

За съвременната наука е безспорна необходимостта от изучаването на детската реч, неопровержимо свидетелство за което е широкото приложение на онтогенетичните данни в качеството им на доказателствен материал при решаването на най-различни проблеми на лингвистиката, а така също и в процесите на търсене и изграждане на адекватни модели на езиковата способност на човека. Показателно за това е направеното от Стефан Младенов още през 30-те години на XX век признание: „Изобщо развитието на детския говор дава материал за осветление и даже за разрешение на току-речи всички главни и второстепенни въпроси на езиковата история, значи не само из областта на фонетиката и морфологията, но и на етимология, словообразуване, лексикология и синтаксис [...]. Който не познава развоя на детския говор, не знае нищо и за езиковата дейност на възрастните“ (Младенов 1934: 30 – 31). Споделянето и разгръщането на подобни идеи не е рядкост нито за българската, нито за световната научна традиция. Достатъчно е само да споменем имената на знакови за хуманитаристиката учени като Я. Б. де Куртене, Л. Блумфийлд, Л. В. Щерба, Р. Якобсон и др. Важността на онтогенетичните данни за лингвистиката сама по себе си обаче не дава обяснение за необходимостта от създаването на корпус с детска реч. Още повече че изучаването на този екзотичен и своеобразен феномен има своята дълга и богата биография, в която обаче винаги е стоял отворен проблемът за надеждността както на самия емпиричен материал, така и на методите за неговото събиране, систематизиране и обработка.

В предложената работа е заложена идеята, че корпусната перспектива би могла да се определи като доминираща в областта на изследванията на езиковата онтогенеза още от времето на Чарлз Дарвин до наши дни. В подкрепа на това биха могли да се приведат множество съществуващи в онтолингвистичната традиция свидетелства за това, че акумулирането на емпирични данни винаги е било, е и ще бъде доминиращо. Тук обаче тези свидетелства ще бъдат представени обобщено и накратко с цел да се открояят спецификациите в хода на своеобразната еволюция в разработването на речевите корпуси, които се екстраполират в адекватен за съответния изследователски период формат.

В периода на първите систематични проучвания върху усвояването на езика се наблюдава предпочитание към събиране на емпиричен материал, т.е. описването на детската реч в нейните конкретни проявления и съхраняването на сведенията за първите думи и изказвания на детето. Така от средата на XIX до средата на XX век е характерна практиката да се водят дневници на хронологичното развитие на детската реч, които имат предимно дескриптивен характер. Философи, естествоизпитатели, лингвисти, психолози правят подробни записки върху речевото развитие на своите деца. Достатъчно е само да споменем имената на някои от тях: Иполит Тен, Чарлз Дарвин, Д. Тидеман, В. Леополд, Грегоар, Я. Б. де Куртене, А. Н. Гвоздев, И. Георгов. Голяма част от тях са публикувани частично или цялостно докъм 30 – 40-те години на XX век. И до днес тези своеобразни „бебешки биографии“ не губят своята значимост. Съвременните учени се връщат непрекъснато към тези данни.

През 30-те години на XX в. в рамките на бихейвиоризма се реализират първите срезови проучвания, в които вече могат да се сравняват образци от много деца на една и съща възраст, а това прави възможно да се прилагат разнообразни статистически методи, да се планират и провеждат експерименти.

С началото на 60-те години на XX век започва епохата на „лонгитудиналните срезови изследвания“, която е своеобразен синтез на методологичните постижения на двата предходни етапа. Документални записи на речеви фрагменти върху магнетофонна лента, които се осъществяват по определен график с предварително назначени времеви интервали, дават възможност да се преодолеят фрагментарността и случайността, присъщи на дневниците и на срезовите данни.

В зависимост от поставената цел отделните методи имат своите предимства и недостатъци. Така например дневниците са много полезни при проучването на онтогенезата на лексикона, но те не са подходящи за получаване на надеждни количествени резултати; срезовите изследвания дават обемна база от данни, но не са в състояние да отчитат достатъчно индивидуалното в езиковото усвояване (най-вече скоростта на това усвояването при различни деца). Лонгитудиналните изследвания дават сравнително точна картина за отделното дете, но събирането на данните отнема много време, а методите за тяхното транскрибиране и обработка показват твърде голямо разнообразие. Това на свой ред прави опирането на един единствен подобен случай твърде неприемливо, а сравняването му с други лонгитудинални данни все по-трудно, тъй като

на практика се оказва, че в отделните корпуси са кодирани специфични индивидуални различия в процеса на усвояването на езика.

С течение на времето и с развоя на техническия прогрес обаче се стига до ново качество на емпиричните продукти и възможностите за тяхната обработка. Картотеките и дневниците са заменени с електронни речеви масиви, трудоемката и изтощаваща работа по регистрирането, транскрипцията и статистическата обработка на данните е осигурена от разнообразни съвременни технически средства и програмни продукти (виж по-подробно Попова 2006: 18 – 26). Това дава основание появата на компютърни системи за натрупване и автоматична обработка на огромни масиви от детските речеви данни да се определи като качествено нов етап в изследванията на езиковата онтогенеза. Именно те през последните няколко десетилетия създават условия за успешното реализиране на мащабни кроссингвистични проекти, посветени на онтогенезата на множество езици. Една от най-популярните компютърни системи е **CHILDES** (**C**hild **L**anguage **D**ATA **E**xchange **S**ystem). Нейното начало се свързва с имената на американските учени Б. Макуини и К. Сноу.

В обобщение на казаното в този параграф може да се направи изводът, че създаването на CHILDES бележи апогея на проследения в резюмиран вид тук еволюционен процес. Типологичното многообразие на включените езикови данни, единният формат за транскрипция, пакетът от програмни ресурси CLAN за автоматична обработка превръщат тази система в една изключително полезна и удобна платформа за изследователска работа (виж по-подробно Попова 2006: 22 – 35). Същевременно може да се добави, че именно оптималните емпирични възможности биха могли да гарантират на всяко едно лингвистично изследване постигането на висока степен на обективност и адекватност на получените резултати, както и да бъдат солидната база за апробация на моделите на езиковата онтогенеза. В контекста на това напълно разбираем е изборът на тази платформа при създаването на български корпус с данни от спонтанна детска реч, който се представя в тази работа.

2. Системата CHILDES – общо представяне

Както вече беше отбелязано по-горе, основна задача на предлаганата работа е да се представи един български компютърен корпус, в който лингвистичните ресурси са транскрибирани и анотирани в термините на системата CHILDES, като при това се маркират и възможностите за приложение на корпусната лингвистика в проучванията на

детската реч, осъществени в светлината на холистичната традиция в съвременната онтолингвистика, за които на свой ред тази платформа осигурява условия за използване на модерни мултимедийни устройства и софтуерни продукти.

Преди да се представи българският корпус, е необходимо да се очертаят рамките на системата CHILDES, в чийто формат са организирани речевите данни на няколко български деца, които са включени в него. Най-важното за изследователите е нейната достъпност. Всъщност става дума за некомерсиална мултимедийна платформа, предоставена в свободна за безплатен достъп зона в интернет (тя е публикувана на адрес: ‹http://childes.psy.cmu.edu›).

Самото название CHILDES, което е абревиатура на Child Language Data Exchange System, директно насочва към това, че става дума за *СИСТЕМА ЗА ОБМЕН НА ДАННИ ПО ДЕТСКА РЕЧ.* Но в действителност идеята за създаването на мащабен международен архив от данни за детския език не е нова. И по-рано е имало няколко индивидуални опита да се споделят данни – например оригиналните записи на Адам, Ева и Сара на Роджър Браун (1973) са били напечатани върху шаблони и преписани на циклостил в множество копия, които са раздадени за ползване на други изследователи, при което по едно главно копие от всеки оригинал се запазва в досиетата на Р. Браун като основен исторически архив. Появата на нова технологична възможност при използването на микрокомпютърните системи на базата на WORD позволява на изследователите да въвеждат данни от записи в компютърни файлове, които след това лесно се размножават, редактират и анализират посредством стандартни обработващи техники. Съхраняването и обменът чрез компютри довежда до промяна в разбирането на самото понятие *архив*. Вместо да е просто хранилище на данни, компютърният архив се оказва един постоянно увеличаващ се набор от данни, обогатяван от всеки, който го използва, защото всеки, който заема нещо от системата, в същото време допринася за нейното разширяване и развитие. Именно в този исторически контекст през 80-те години на XX век се появява CHILDES. Тя представлява динамична система за обмен на данни от детската реч и се поддържа от екип учени начело с Брайън Макуини и Кетрин Сноу (от университета Карнеги Мелън в САЩ).

Свидетелство за изключителната жизненост и устойчивост на CHILDES са както нейното над 30-годишното съществуване и непрекъснато нарастващ брой на участниците (над 3200), така също и нейното интегриране във възникналата и утвърдила се през първите го-

дини на второто хилядолетие широкомащабна платформа за изучаване на комуникацията – TalkBank (<http://www.talkbank.org/>).

Базата от данни на CHILDES съдържа голям обем от сведения за усвояване на множество езици, като например английски, африкаанс, датски, холандски, френски, немски, иврит, унгарски, италиански, полски, испански, турски, хърватски, руски, словенски, сръбски и др., като техният брой непрекъснато расте. В базата данни има и специален раздел за аномалиите в езиковото развитие и за усвояването на втори език. Файловете са предоставени за свободен достъп в интернет на адрес: <http://childes.psy.cmu.edu/data/>.

Някои от данните на системата CHILDES са представени от транскрипти, които са свързани с аудио- и видеофайлове. Тези мултимедийни файлове създават възможности на изследователя за систематично проучване на нови аспекти на детския език. Например, въпреки че някои от стандартните CHAT транскрипти включват информация за интонацията и илокутивната сила, то е невъзможно систематичното изследване на тези аспекти без осигурена постоянна връзка между транскриптите и аудиозаписите. Що се отнася за видеофайловете, в това отношение би могло да се отбележи значението на тяхната връзка със съответстващите им транскрипти за създаването на оптимални възможности за изучаването на невербална комуникация и влиянието на прагматичния контекст върху речта на детето.

Освен базата данни CHILDES предоставя на изследователите и пакет със специализирани програми CLAN, чрез които е възможно да се осъществява различен тип анализ на въведените диалози (фонетичен, морфологичен, синтактичен) и коментарите към тях. В този смисъл CLAN дава възможност автоматично да се получат най-разнообразни статистически и съдържателни резултати от транскрибираните и кодирани данни, като например за честотата на думите, за лексикалното разнообразие и съчетаемост, за специфичните потребителски думи и форми (например детските езикови грешки като специфични отклонения от нормата на съответния език: единиците на т.нар. BABY TALK, ономатопеи, свръхгенерализации, детски и семейни оказионализми) и т.н.

Несъмнена е ползата от автоматизираната компютризирана система за обмен на езикови данни CHILDES. Причините за нейното разработване са очевидни за всеки, който е създавал и анализирал записи. Една такава система дава възможност да се осигури по-голяма научна прецизност при събирането, транскрибирането и кодирането на данните, а също така да се автоматизира анализът на големи количес-

тва разговорен материал, което разширява значимо емпиричната база, върху която се строят новите теории. Всичко това допринася за постигане на ново качество на научния продукт. Безспорно свидетелство за предимствата на CHILDES са над 3200-та публикувани работи върху различни езици (над 33 езика), базирани върху използването на тази система (Child Language Bibliographies е публикувана на адрес: <http://childes.talkbank.org/bibs/>).

Системата CHILDES е особено необходима днес, когато се осъществяват мащабни интегративни изследвания на детската реч в рамките на международни научни проекти (като напр. крослингвистичните проекти Pre- and Protomorphology in Language Acquisition[1]; Syntaktische Konsequenzen des Morphologieerwerbs[2]; Erwerb sprachlicher Markierungen zur Differenzierung von ±Begrenztheit[3;] Spracherwerb: Acquisition and Disambiguation of Intersentential Pronominal Reference[4] и др.). В този контекст именно универсалният модел за представянето и за анализа на данните, съдържащ се в CLAN, дава на учените, изучаващи по няколко десетки езика, резултатно и надеждно да осъществяват сравнително-типологични изследвания и на тази база да се строят солидни модерни теории.

3. Електронен корпус на българска детска реч – описание

Създаденият в Лабораторията по приложна лингвистика на Шуменския университет Електронен корпус на българска детска реч е резултат на дългогодишна работа, а именно от 1990 г. до наши дни. Той е разработен в термините на CHILDES и включва два типа речеви ресурси: **ПОДКОРПУС А**, включващ спонтанна реч на 4 деца в ранна възраст (от 1 до 3 години), и **ПОДКОРПУС Б**, включващ разкази по серия картинки на 90 деца в предучилищна възраст (от 3 до 6 години). Този емпиричен материал все още не е включен в общата банка на системата CHILDES, тъй като и в момента продължава разширяването му с нови данни, а също така се работи и по подготовката за включване на натрупаните необработени масиви. В този смисъл условно може да се говори и за **ПОДКОРПУС В,** който е в режим *UNDER CONSTRUCTION* (т.е. намиращ се в процес на изграждане).

[1] Виж <http://www.oeaw.ac.at/ ling/kimo/international_prepro.html>
[2] Виж <http://www.zas.gwz-berlin.de/fileadmin/material/ jahresberichte/jb2000.pdf>
[3] Виж <http://www.zas.gwz-berlin.de/fileadmin/ material/jahresberichte/ jb2003.pdf>
[4] Виж <http://www.zas.gwz-berlin.de/fileadmin/material/jahres berichte/jb2007.pdf>

Досега събраните и подготвени за компютърния архив CHILDES данни включват аудиозаписите на спонтанна реч на четири български деца (преобразувани в компютърни WAV файлове), които са транскрибирани и кодирани в CHAT формат (в отделните транскрипти децата са обозначени съответно като *TEF, *ALE, *BOG, *IVE). Това създава необходимите условия за успешното използване на пакета със специализирани програми CLAN и става възможно да се осъществяват различни анализи на въведените диалози и коментарите към тях. При това всеки потребител има свобода според целите, които си поставя в дадено конкретно изследване, и сам да си създава допълнителни редове с коментари, които са му необходими. Коментарите могат да бъдат най-различни – фонетични, морфологични, ситуационни, авторски, съответно представени в CHAT файла като специални линии, а именно: **%pho**, **%mor**, **%sit**, **%com** и др. В илюстрация на казаното е предложен откъс от реален *.cha файл от електронен корпус, в който при транскрипцията са въведени два типа коментарни редове – морфологични (%mor) и ситуационни (%sit). При това ситуационните коментари се появяват в случаите, когато определено изказване се оказва неясно извън контекста, докато морфологичните съпътстват всяка една реплика на изследваното българско момиче Стефани, тъй като съответните емпирични данни са подготвени за автоматичен анализ на ранната глаголна онтогенеза с помощта на програмите от пакета CLAN. Срв.:

```
@Begin
@Participants: TEF Stefani Target_CHILD, VEL Velka Experimenter, BAB Rosica Grandmother
@Birth of TEF:   29-NOV-2000
@Age of TEF      1;08.0
@File name:      St_290702.cha
@Tape Location: Cassete 1, Side B
@Date:  29-JUL-2002
@Situation:      at home
*VEL:   Koj e tozi?
%sit:    posochva koteto, koeto jade
*TEF:   Papa.
%mor:   V|papam&IPFV:TRANS-PRES:3S
*VEL:   Koj papa?
*TEF:   Mau.
%mor:   ONOM|mjau
*VEL:   Kakvo papa mau?
*TEF:   Papa.
%mor:   V|papam&IPFV:TRANS-PRES:3S
*VEL:   Papa mau?
%sit:    utochnjava
*TEF:   Mau-mau!
%mor:   ONOM|mjau-mjau
%sit:    TEF dyrpa koteto
*BAB:   Leko, babo, shche plache koteto!
*TEF:   Pachi [:plache].
```

```
%mor:   V|placha&IPFV:INTRANS-PRES:3S
*TEF:   Papa.
%mor:   V|papam&IPFV:TRANS-PRES:3S
*BAB:   Neka da papa mau!
*BAB:   Toj e gladen.
*BAB:   Daj mu da jade!
*TEF:   De [:jade].
%mor:   V|jam&IPFV:TRANS-PRES:3S
%sit:   povtarja
*BAB:   Kakvo shche pravish sega?
*TEF:   Pija.
%mor:   V|pija&IPFV:TRANS-PRES:1S
*VEL:   A lelja kakvo shche pie?
*TEF:   Baba.
%mor:   N|baba&FEM-SG
%sit:   TEF nabljudava kak VEL i BAB pijat kafe
*VEL:   I baba pie kafe.
*TEF:   Pij [:pie].
%mor:   V|pija&IPFV:TRANS-PRES:3S
%sit:   povtarja
.....
@End
```

Информацията, която носят тези допълнителни коментарни редове, е особено важна при изследванията както на речта на малките деца, тъй като тя изобилства с отклонения от нормата и е силно ситуативна, така и на особеностите на усвояване на втори език (L_2), на интеракцията между възрастни и деца, на възстановяването на езика при афазия и т.н. Всяко разширяване на заложените в транскриптите данни довежда до възможност както за тяхната приложимост в различни нови области, така и за използване на повече от програмите на CLAN, което на практика ги прави полезни за доста широк кръг специалисти.

Подкорпус А

Досега в рамките на **ПОДКОРПУС А** събраните и подготвени за компютърната база CHILDES данни включват аудиозаписите на спонтанна реч на четири български деца (преобразувани в компютърни WAV файлове), които са транскрибирани и кодирани в CHAT формат с добавени коментарни линии от рубриката „%sit“. Това са данните на четири български деца: Александра (ALE), Стефани (TEF), Богомила (BOG) и Ивелин (IVE).

Транскрибирането е осъществено в Sonic Mode, като веднага трябва да се отбележи, че този метод изисква повече време, но за сметка на това е изключително прецизен (вж. по-подробно Макуини, Вагнер 2010: 4 – 5), срв. Фигура 1.

Фигура 1. *Аудиотранскрипция*

В основата на базата данни са заложени 33 часа записи (дигитализирани и съхранени в 32 *.wav файлове) и транскрипти в 355 страници. При създаването на *.cha файлове данните от някои краткосрочни и близки по време файлове бяха обобщени в един документ. Корпусът от данните на четирите деца е представен в 30 файла в CHAT формат.

Наблюдаваните деца са родени и живеят в град Шумен, Североизточна България. Те са записвани в обичайни ситуации (игра, обличане, хранене, приспиване, разглеждане на книжки с картинки и т.н.) в процеса на ежедневното им общуване с най-близките. Всички лица, регистрирани в базата данни като участници в диалозите, са монолингви, носители на български език. Възрастните от обкръжението на децата са с добро ниво на образование (средно гимназиално и университетско). Аудиозаписите на три от децата (ALE, TEF, IVE) са направени от авторката на тази статия, която е и майка на едно от децата (ALE), а на BOG – от майката на детето (също лингвистка). Транскрипцията и кодирането на материала са изцяло дело на авторката на настоящата статия.

Подкорпус Б

В рамките на **подкорпус Б** попадат **разказите** на деца от 3- до 6-годишна възраст от Шумен и Варна. Те бяха записвани с диктофон, след което съответните аудиозаписи бяха преобразувани в компютър-

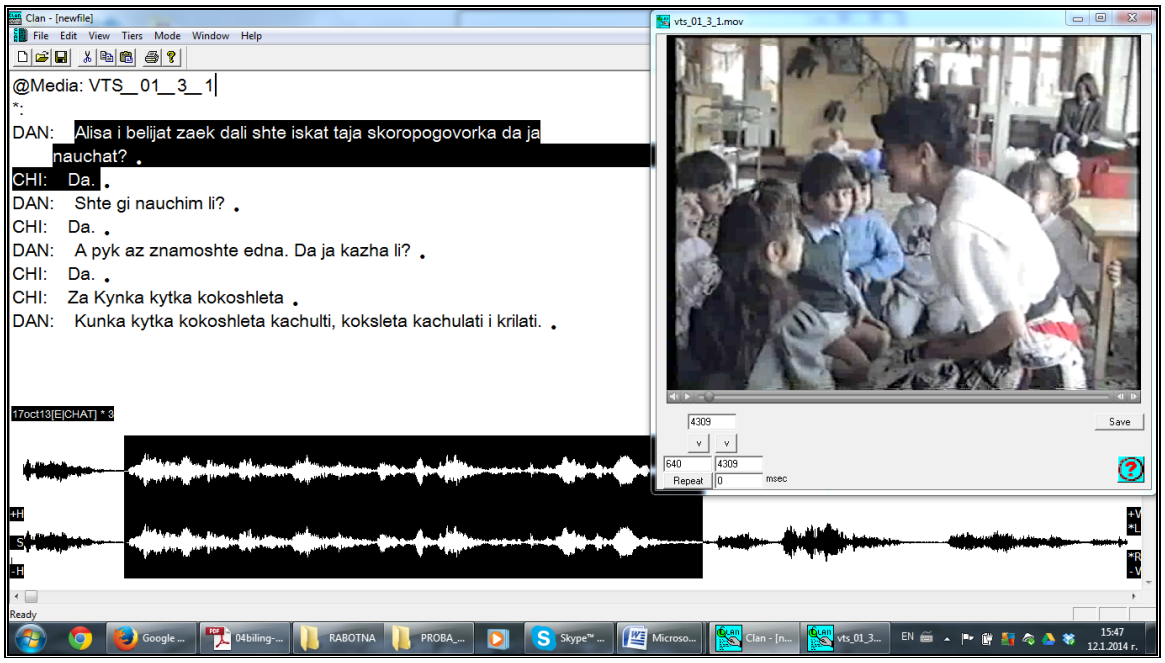
ни WAV файлове, които са транскрибирани (също както в КОРПУС А **в Sonic Mode**) и кодирани в CHAT формат на системата CHILDES. В базата данни са заложени 3 часа записи (дигитализирани и съхранени в 30 *wav файла) и 60 транскриптаи в 62 страници. Записите са направени от 3 учителки в детски градини на град Шумен, а транскрипциите и анотирането – от авторката на тази публикация.

Описаните дотук подкорпуси (**А** и **В**) с данни от спонтанната речева продукция на български деца в ранна и предучилищна възраст беше финансово и методически подкрепена от Центъра по общо езикознание в Берлин (ZAS Berlin). Тяхната приложимост и надеждност отчасти вече е апробирана в рамките на дискусиите и сравнителните анализи на българския с други езици (немски и руски), осъществявани в рамките на крослингвистичната програма за изследване на ранното усвояване на аспекта (срвн. Кюнаст, Попова, Попов 2004; Битнер, Гагарина, Попова, Кюнаст 2005). Същевременно корпусът заляга като емпирична база и на множество частни изследвания върху различни страни на ранната онтогенеза на българската граматика (вж.: Попова, Попов 2007; Попова 2010; Попова 2011 и др.), а така също и на проучвания на детския език в предучилищна възраст, които са все още в работен режим.

Подкорпус В

ПОДКОРПУС В, както вече беше отбелязано, е все още в режим *UNDER CONSTRUCTION*. Така подготвяната българска база данни обхваща също и видеоданни, които биха могли да бъдат отнесени към рубриката **ClassTalk** (*говорене в класната стая*) на системата TalkBank, които се интегрират естествено с вече подготвените за CHILDES корпуси. Те представят рубриката **Classroom interactions** (*взаимодействия в класната стая*) и обхващат няколко учебни занятия в детска градина. Към момента се намират в етап на работен вариант, тъй като все още се транскрибират. (Транскрипцията на видеофайлове следва същите основни принципи, които се използват и при аудиотранскрипцията. **Фигура 2** илюстрира как се транскрибират видеофайлове в CLAN – виж по-подробно Макуини 2007.) Тези корпуси биха могли да се използват както за изследване на спецификата на речевата интеракция между учителя и децата в рамките на учебно занятие в детската градина, така и като нагледен материал в процеса на обучение на студенти педагози.

Фигура 2. *Видеотранскрипция с отворен звуков панел*



Българските данни все още не са публикувани в мрежата на CHILDES, както вече беше казано по-горе, но въпреки това представяният корпус вече е използван неколкократно в различни публикации и международни проекти като емпирична база за апробация на когнитивни модели на речевата интеракция. Очакванията за бъдещото му разширяване и оптимизиране са свързани със създаването на паралелни корпуси, което би било изключително важно за съпоставителните изследвания на различни езици и култури.

В Лабораторията по приложна лингвистика на Шуменския университет продължава натрупването на лингвистичен материал и се работи по подготовката на корпусите с необработените данни. (Така напр. в режим на подготовка се намира и корпусът с данни от свободен асоциативен експеримент с 50 български деца в предучилищна възраст, сред които има и монолингви, и билингви (български – турски и български – руски). На настоящия етап напълно готова е дигиталната аудиобаза и предстои осъществяването на транскрибирането на речевия материал.)

Не на последно място трябва да се отбележи още и това, че унифицираната удобна техника за анотиране на екстралингвистичните данни, които съпътстват речта на наблюдаваните лица, както и непрекъснатият режим на връзка между транскриптите и съответните аудио- и видеофайлове създават възможности и перспективи не само за изследване на всички аспекти на речевата интеракция, но и на общу-

ването като цяло от гледна точка и на останалите науки. В този смисъл чрез CHILDES учени от различни сфери на хуманитаристиката биха могли да реализират успешно своите търсения, както и да обединят усилията си в интердисциплинарни изследователски проекти. В логиката на тези думи в перспектива би могло да се очаква системата за обмен на данни CHILDES да се превърне в една от най-успешните професионални ONLINE мрежи за хуманитаристи, което на свой ред би осигурило възможности за солидни и модерни междудисциплинарни изследвания.

Заключение

Представеният Електронен корпус с българска детска реч е само миниатюрен фрагмент от една многоезична виртуална мозайка, която непрекъснато се разширява и обогатява. В предложената работа беше направен опит той да бъде представен в контекста на двете мощни системи – CHILDES и TalkBank, които със своята отвореност и рационалност се налагат като водещи в процесите на кооперация и глобализация в хуманитаристиката като цяло. А това е гаранция както за широка социална валидност на резултатите от изследванията, базирани на техните корпуси, така и за интегрирането им в актуалните работни програми за създаване на инфраструктури за обмен на езикови данни и технологии, целящи преодоляване на сегашната разпокъсаност на научното пространство. В този смисъл приносът на представения тук български корпус би могъл да се разчете в неговата полезност както за лингвистиката и другите науки, така и за обществото. Свидетелство за това е включването му в програмата за изграждането на BG-CLARIN[5], който е част от CLARIN (Common Language Resources and Technology Infrastructure), имащ за цел създаването на Европейската инфраструктурна мрежа.

Литература

Битнер и кол. 2005: Bittner, D., Gagarina, N., Popova, V., Kühnast, M. Aspect before Tense in the acquisition of Russian, Bulgarian, and German. // V. Solovyev, V. Polyakov (Eds.). *Text Processing and cognitive Technologies*. Moskow: Ucheba, 263 – 272.

[5] Виж подробно за това в „Национална пътна карта за научна инфраструктура“ на адрес: <http://www.strategy.bg/StrategicDocuments/View.aspx?lang=bg-BG&Id=624>.

Браун 1973: Brown, R. *A first language: The early stages.* Cambridge, MA: Harvard University Press, 1973.

Кюнаст, Попова, Попов 2004: Kühnast, M., Popova, V., Popov, D. Erwerb der Aspektmarkierung im Bulgarischen. // N. Gagarina, D. Bittner (Eds.). *ZAS-Paper in Linguistics 33, 2004. Studies on the development of grammar in German, Russian and Bulgarian*, 2004, 63 – 87 <http://alphalinguistica.sns.it/Riviste/ZAS/ 33_2004.pdf>.

Макуини 2007: MacWhinney, B. Opening up video databases to collaborative commentary. // R. Goldman, R. Pea, B. Barron, Sh. Derry (eds.). *Video research in the learning sciences*. Mahwah: Lawrence Erlbaum Associates, 2007, 537 – 546.

Макуини, **Вагнер 2010:** MacWhinney, B., Wagner, J. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. // *Gesprächsforschung* – Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837) Ausgabe 11 (2010), 154 – 173 <*www.gespraechsforschung-ozs.de*>.

Младенов 1934: Младенов, Ст. Няколко езикословни въпроси у проф. д-р Ив. А. Георгов в работите му за развоя на детския говор. // *Год. СУ. ИФФ*. кн. XXX, 1934, 1 – 35.

Попова, Попов 2007: Popova, V., Popov, D. The emergence of verb grammar in two Bulgarian-speaking children. // V. Solovyev, V. Polyakov (Eds.). *Text Processing and cognitive Technologies*. Moskow, 2007, 236 – 248.

Попова 2010: Попова, В. Корпусно изследване на граматичната метаморфоза на ранния детски език. // *Език, култура, идентичност*. Велико Търново: Фабер, 2010, 101 – 115.

Попова 2011: Попова, В. Ролята на ономатопеите в ранната глаголна онтогенеза. // *Litera et Lingua*. *Пролет 2011* (Електронно списание на Факултета по славянски филологии на СУ), София, 2011 <http://slav.uni-sofia.bg/lilijournal/index.php/bg/issues/spring2011>.