Български език, Приложение / Bulgarian Language, Supplement, 71 (2024), 328–346
Print ISSN: 0005-4283, Online ISSN: 2603-3372
doi: 10.47810/BL.71.24.PR.21

Колекция с българска детска реч в термините на корпусната лингвистика

Велка Попова
Шуменски университет „Епископ Константин Преславски“
v.popova@shu.bg

Предлаганата работа е фокусирана върху възможностите за приложение на корпусната лингвистика при изследване на речевата интеракция *дете – възрастен*, като по-специално вниманието е насочено към автоматизираната компютризирана система за обмен на езикови данни CHILDES, предоставена за свободен достъп в интернет – https://childes.talkbank.org.

Основна цел на доклада е да се представи първият български компютърен корпус *Bulgarian LabLing Corpus* (https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html), в който лингвистичните ресурси са транскрибирани и анотирани в термините на тази платформа, като при това се прави опит да се открият ползите от нея. Те намират израз в това, че CHILDES осигурява условия за по-голяма прецизност при събирането, транскрибирането и кодирането на данните, а също така и в автоматизирането на анализа на големи масиви разговорен материал. Това разширява значително емпиричната база, върху която се строят новите теории.

Ключови думи: корпусна лингвистика; CHILDES; детска реч; български език

A Collection of Bulgarian Child Language in Corpus Linguistics Terms

Velka Popova
Konstantin Preslavsky University of Shumen
v.popova@shu.bg

The proposed article focuses on the application possibilities of corpus linguistics to the research of the *child-adult* speech interaction, and more specifically – the CHILDES automated computerized system for language data exchange, freely accessible online at https://childes.talkbank.org.

The main aim of the report is to present the first Bulgarian corpus, *Bulgarian LabLing Corpus* (https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html), where the linguistic resources are transcribed and annotated according to the terms of the platform, as well as to point out the advantages of the latter. Those include the conditions provided by CHILDES for improved precision in the collecting, transcribing and coding of the

data, along with the automation of the analysis of large quantities of spoken material, which considerably enlarges the empirical basis on which new theories are built.

Keywords: corpus linguistics; CHILDES; child language; Bulgarian language

Въведение

За съвременната наука е безспорна необходимостта от изучаването на детската реч. Неопровержимо свидетелство за това е широкото приложение на онтогенетичните данни в качеството им на доказателствен материал при решаването на най-различни проблеми на лингвистиката. 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/Mladenov 1934: 30–31).

Важността на онтогенетичните данни за лингвистиката сама по себе си обаче не дава обяснение за необходимостта от създаването на корпус с детска реч. Изучаването на този 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. Същевременно с налагането на антропоцентричната парадигма в съвременната хуманитаристика настъпват съществени промени в цялостната концепция за изследване на детската реч, като във фокус вече не е изолираното проучване на нейните единици и особености, а самото говорещо дете. От тази гледна точка е напълно обяснимо обръщането на учените към холистичната парадигма, в която се поддържа тезата за неделимо преплитане на общи езикови и когнитивни принципи и правила. Това на свой ред измества интереса от традиционното лингвистично обяснение на езиковите правила и принципи към търсенето на съответните отговори от позициите на когнитивната лингвистика. Като ключов проблем на изследванията вече се откроява проучването на природата на езиковата познавателна система и достъпът до нейния процесор.

В контекста на казаното възниква въпросът за емпиричното осигуряване на когнитивната холистична парадигма. При това възможните решения могат да бъдат намерени в областта на съвременната корпусна лингвистика, тъй като именно в нея се продуцират бази данни, солидни не само по обем, но и по отношение на възможностите за комплексно изучаване на речта.

Именно това категорично наложи като изследователска парадигма през последните десетилетия корпусните проучвания, осъществявани в светлината на холистичната традиция в езикознанието. А благодарение и на „постиженията на съвременните информационни технологии и създадените

мултимедийни програмни пакети, предназначени за постигането на различни цели в отраслите на приложната лингвистика, за съвременния изследовател стана възможно илюстрирането и визуализацията на резултатите да бъде осъществявано комплексно и цялостно. Затова допринасят и натрупванията в младата мултимодална лингвистика, представляваща едно от направленията за едновременно многоизмерно представяне на езиковите факти и явления в рамките на холистичното езикознание“ (Попов/Popov 2016: 16).

В логиката на казаното става ясно, че неслучайно настоящата работа е посветена на корпусния подход в изследването на детската реч, като неговите предимства са представени по примера на някои съвременни, добре работещи изследователски платформи като CHILDES (https://childes.Talkbank.org/) и TalkBank (https://www.talkbank.org/). И по-конкретно във фокус се оказва първият електронен корпус на българска детска реч (Bulgarian LabLing Corpus). При това е направен опит да се откроят 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 опасността от появата на нови митове в съвременните научни търсения.

В този ред на мисли напълно естествено възниква въпросът 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 следване на някаква мода? От какъв род (не)достатъчност „страдат“ познатите и добре работещи традиционни модели? В отговор на тези въпроси ще бъде предложен кратък хронологичен преглед 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. Корпусната перспектива в онтолингвистиката

Детската реч е екзотичен и своеобразен феномен, който има своята дълга и богата биография, в която обаче винаги е стоял проблемът за надеждността както на самия емпиричен материал, така и на методите за неговото събиране, систематизиране и обработка. В този смисъл предложената работа е вдъновена от идеята, че корпусната перспектива би могла да се определи като доминираща в областта на изследванията на езиковата онтогенеза от времето на Чарлз Дарвин до днес. 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 доминиращо и ще доминира и в бъдеще.

Тук обаче тези свидетелства ще бъдат представени обобщено и накратко с цел да се открият спецификациите в хода на своеобразната еволюция в разработването на речевите корпуси, които се екстраполират в адекватен за съответния изследователски период формат. Във връзка с това в историята на проучванията на детската реч биха могли да се обособят няколко последователни етапа:

- Етап на т.нар. *бебешки биографии* от средата на XIX до средата на XX век, които се свързват с практиката да се водят дневници на хронологичното развитие на детската реч. Те имат предимно дескриптивен характер. (Подробни записки върху речевото развитие на своите деца правят философи, естественици, лингвисти, психолози, като сред тях са Иполит Тен, Чарлз Дарвин, Д. Тидеман, В. Леополд, Ян Бодуен де Куртене, А. Н. Гвоздев, И. Георгов.)
- Етап на появата на първите срезови проучвания през 30-те години на XX в., в които вече могат да се сравняват образци от много деца в една и съща възраст. Това прави възможно да се прилагат разнообразни статистически методи, да се планират и провеждат експерименти.
- Етап на лонгитудните срезови проучвания, чието начало се свързва с 60-те години на XX в., в който се наблюдава своеобразен синтез на методологичните достижения на двата предходни етапа, доколкото те създават възможност да се преодолее фрагментарността и случайността, присъща за дневниците и на срезовите данни.
- Етап на създаването на компютърни системи за натрупване и автоматична обработка на огромни масиви от детските речеви данни, даващ възможности да се преодолеят недостатъците на методите от предходните етапи, а именно: въпреки че се отчита полезността на дневниците за проучване на онтогенезата на лексикона, те не са подходящи за получаване на надеждни количествени резултати; срезовите изследвания на свой ред дават обемна база от данни, но не са в състояние да покажат достатъчно индивидуалното в езиковото усвояване; лонгитудните изследвания дават сравнително точна картина за отделното дете, но трудоемкостта на събирането и обработката на данните и различните методи на транскрибиране прави сравнението с други лонгитудни изследвания извънредно сложно, а проучването само на един отделен корпус би могло да бъде подвеждащо поради предопределени индивидуални различия в процеса на усвояването.

Последният етап в изследванията на лингвистичната онтогенеза може да се определи като качествено нов, тъй като именно изгражданитее в неговия обхват компютърни системи за натрупване и автоматична обработка на огромни масиви от данни с детска реч през последните десетилетия създават условия за успешното реализиране на мащабни крослингвистични проекти, в които са включени множество езици. При това лингвистичните ресурси, организирани като корпуси, все по-широко се използват за целите на

моделирането на езика и на речевото поведение на неговите носители въпреки факта, че създаването и поддържането на компютърни корпуси е едно извънредно трудоемко и скъпо струващо начинание.

С появата на модерните технологии се налага по-ефективен стандарт за тяхното представяне и обработка. По този начин става възможно разширяването на обхвата на отделния корпус до милиони езикови единици, като същевременно се оптимизират и опциите за тяхното анотиране, унифициране, стандартизиране и многократно използване.

Така с течение на времето се стига до замяна на картотеките и дневниците с електронни речеви масиви. Трудоемката и изтощаваща работа по регистрирането, транскрипцията и статистическата обработка на данните е подкрепена от разнообразни технически средства и програмни продукти. А създаването на системи като CHILDES и TalkBank бележи апогея на проследения в резюмиран вид еволюционен процес. Типологичното многообразие на включените езикови данни, единният формат за транскрипция, пакетът от програмни ресурси CLAN за автоматична обработка превръщат тази система в изключително полезна и удобна платформа за изследователска работа. 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, които да бъдат и солидната база за апробация на моделите на езиковата онтогенеза. В контекста на това напълно разбираем е изборът на платформата CHILDES при създаването на българския корпус с данни от детска реч, който се представя в тази работа.

В следващата част от изложението ще бъде направен опит да се даде представа за новото качество на изследванията от последния етап с оглед на българската онтолингвистика. Така съвсем естествено във фокус попада автоматичната компютърна система CHILDES като една от най-популярните платформи за обмен на данни за детска реч и в частност един от нейните корпуси, а именно *Bulgarian LabLing Corpus*. Основната цел е да се демонстрира, че този формат на представяната българската колекция с данни от детска реч не е плод на мимолетна лингвистична мода, а по-скоро възможност за осигуряване на надежден емпиричен материал за създаване на адекватни онтогенетични модели, както и за постигането на качествено нов стандарт на работа за съвременния изследовател.

Същевременно обръщането към CHILDES е предопределено и от факта, че по примера на тази система могат да бъдат откроени максимално възможностите за приложение на корпусната лингвистика при изследването на речевата интеракция *дете – възрастен.* Адекватното ѝ разчитане би могло да даде отговори на множество все още отворени въпроси, свързани с езиковата онтогенеза и усвояването на българския език в зората на човешкия живот.

2. Малък екскурс. Някои терминологични уточнения

Преди да се пристъпи към представянето на българския CHILDES корпус, е необходимо да се уточнят някои терминологични спецификации на названието *детска реч*. Веднага трябва да се отбележи, че то попада в мрежите на изключителния плурализъм в съвременната лингвистична терминология, като при това се асоциира с три различни понятия в контекста на *Многостепенния модел на българската реч* (Попов/Popov 2005), в който на свой ред е вградена трихотомията *стандарт – субстандарт – нонстандарт* (Кёстер-Тома/Kjoster-Toma 1993). В резултат от направената своеобразна апробация на *Модела* при етаблирането на концептите **детска реч$_1$** (речта на самото дете в процеса на общуване с възрастни или с други деца), **детска реч$_2$** (речта на възрастния към детето) и **детска реч$_3$** (стилизирано говорене на възрастния „по детски“ в общуване с възрастни те се оказаха съотнесени съответно към *субстандарта в процеса на усвояването му*, към *субстандарта* и към *нонстандарта* (вж. Попов, Попова/Popov, Popova 2021).

Доколкото предлаганата работа е фокусирана върху интеракцията *възрастен – дете,* в следващото изложение вниманието ще бъде насочено към двата субстандартни феномена, а именно – **детска реч$_1$** и **детска реч$_2$**. Тяхното адекватно разбиране би довело до създаването на работещи модели на езиковата онтогенеза, доколкото специфичният регистър за общуване на възрастните с децата съдържа в себе си някои важни предпоставки за успешното усвояване на езика в ранна възраст, тъй като именно в това общуване системата на езика цел е представена във вид на програма минимум на входа. Така в хода на общуването с децата възрастните доставят входни езикови данни, сред които ясно се открояват т.нар. *позитивни* и *негативни данни* (вж. Стоянова/Stoyanova 2022).

В реда на тези мисли за значимостта на т.нар. *Baby talk* за овладяването на езика от детето естествено се появява идеята за необходимостта от проучването не само на *речта на децата*, а и на тази социална вариация на езика, известна още като *регистър на възрастните към децата*. Във връзка с това в последните няколко десетилетия се акумулираха научни приноси, сред които като особено ценни се откроиха лонгитудните изследвания, създаващи условия да се улови адекватно логиката на езиковия развой. Несъмнено това беше съпроводено и от търсения на подходящи формати на речевите колекции, които да дават оптимален достъп както до речта на децата, така и до речта на възрастните към децата. Именно затова в настоящата работа обект на внимание е оптималната система за диференцирано извличане на данните на отделните изследвани лица CHILDES.

Същевременно CHILDES е максимално отворена система, в която освен лонгитудни данни могат да бъдат представени и данни от срезови проучвания, като например колекции от наративни образци, извлечени от множество деца в една и съща възраст. Именно подобно разнообразие от данни ха-

рактеризира и представяния тук първи български CHILDES корпус (*Bulgarian LabLing Corpus*).

3. Системата CHILDES – кратко представяне

Както вече беше отбелязано по-горе, основна задача на настоящата работа е да се представи първият български компютърен корпус, в който лингвистичните ресурси са транскрибирани и анотирани в термините на системата CHILDES. Преди да се представи българският корпус, е необходимо да се очертаят рамките на системата CHILDES, в чийто формат са организирани речевите данни на изследваните български деца, които са включени в него. Най-важното за изследователите е нейната достъпност. Всъщност става дума за некомерсиална мултимедийна платформа, предоставена свободно в интернет.

CHILDES представлява система за обмен на данни за детска реч, но идеята за създаването на мащабен международен архив от данни за детския език се появява преди това. Вече е имало няколко по-ранни опита да се споделят данни. Например колекцията на Р. Браун (Brown 1973) с оригиналните записи на Адам, Ева и Сара, които са били напечатани върху шаблони и преписани на циклостил в множество копия. Те са били 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 като динамична система за обмен на данни от детската реч.

Базата от данни на CHILDES съдържа голям обем от сведения за усвояване на множество езици, като в нея има специален раздел за аномалиите в езиковото развитие и за усвояването на втори език.

Несъмнена е ползата от тази автоматизирана компютъризирана система за обмен на езикови данни. Причините за разработването на CHILDES са очевидни за всеки, който е записвал 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 количества разговорен материал, което раз-

ширява значително емпиричната база, върху която се строят новите теории. Системата CHILDES е особено необходима днес, когато се осъществяват мащабни интегративни изследвания на детската реч в рамките на международни научни проекти.

4. *Bulgarian LabLing Corpus* на платформата CHILDES

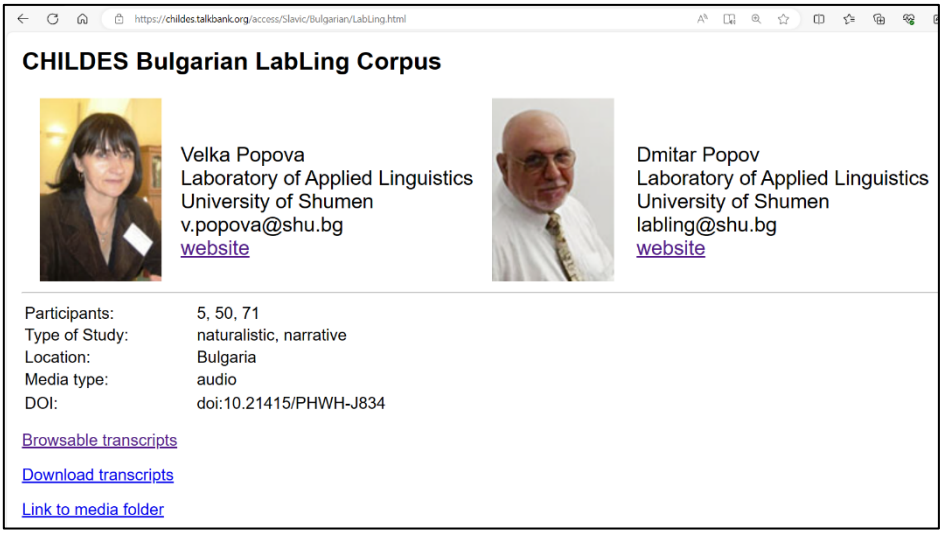
Базата от данни на CHILDES съдържа голям обем от сведения за усвояване на множество езици от различни езикови семейства, като техният брой непрекъснато расте. През есента на 2020 г. в *Славянската колекция* на платформата CHILDES се появи ново попълнение, а именно *Bulgarian LabLing Corpus* в пилотната си версия, която беше доразвита и разширена до финалния си формат през 2023 г. (вж. фигура 1).

CHILDES | Slavic Corpora

Corpus	Age Range	N	Media	Comments
Bulgarian				
LabLing	1-5	5, 50, 71	some audio	5 longitudinal, 50 and 71 narrative
Croatian				
Kovacevic	1;3-2;8 1;10-2;11 0;10-3;2	3	audio	Two girls and a boy learning Croatian in Zagreb
MAIN	5-63	143	audio	MAIN protocol
Czech				
Chromá	1;5-3;5	7	audio	children learning Czech in Prague
Polish				
Szuman	1;5-7;9	10	-	Diary data collected by Szuman and his students and computerized by Magdalena Smoczynska
WeistJarosz	1;7-2;6	3	audio	in ***PhonBank***
Polish-CDS	adults	various	-	child-directed speech from various corpora
Russian				
Protassova	1;6-2;10	1	-	Longitudinal study of a child learning Russian
Tanja	2;5-2;11	1	-	A child learning Russian in a monolingual environment in the United States
Serbian				
SCECL	1;6-4;0	8	audio	Recordings in homes with many people included
Slovenian				
Zagar	5;0	20	-	Arguments patterns in kindergarten children

Фигура 1. Славянски корпуси, публикувани на платформата CHILDES

Bulgarian LabLing Corpus е предоставен на свободен достъпен на адрес: https:// childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html



Фигура 2. Българският корпус с детска реч на CHILDES – Bulgarian LabLing Corpus

Представяният тук корпус с българска детска реч е резултат от дългогодишния труд на изследователи от LABLING. Данните на корпуса са транскрибирани в унифицирания формат CHAT на системата CHILDES (Mac Whinney 2000), което ги прави съотносими с корпусите на други езици от платформата. А пакетът от програми CLAN позволява да се направи различен тип анализ на въведените диалози (фонетичен, морфологичен, синтактичен) и коментарите към тях. В този смисъл CLAN осигурява възможност автоматично да се получат най-разнообразни статистически и съдържателни резултати от транскрибираните и кодирани данни за честотата на думите, за лексикалното разнообразие и съчетаемост, за употребените в съответната речева сесия специфични думи и форми (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.

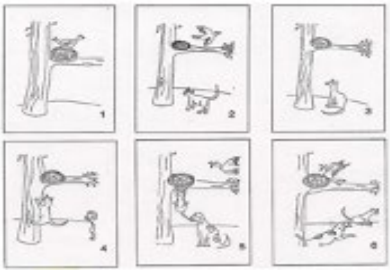
В структурата на Bulgarian LabLing Corpus са обособени два подкорпуса. Единият – лонгитудинален, а другият – кроссекционен. Това са: Корпус А с лонгитудните данни на 5 български момичета в ранна възраст; Корпус Б с 233 наратива на 121 деца в предучилищна възраст, който може да се определи като кроссекционен, доколкото изследваните лица допълнително са разделени по възраст в три групи.

Корпус А обхваща транскрибираните лонгитудни данни от 5 български момичета – ALE (във възрастта 1;01.29–2;04.09), TEF (във възрастта 1;03.11–

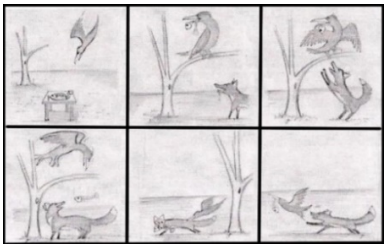
2;05.25), BOG (във възрастта 2;01.09–2;04.11), ELI (във възрастта 1;01.07–1;07.22), SIM (във възрастта 1;04.14–3;01.00). Изследваните деца са родени и живеят в Североизточна България. Те са записвани в обичайни ситуации (игра, обличане, хранене, приспиване, разглеждане на книжки с картинки и т.н.) в 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).

Корпус Б съдържа два сегмента, а именно *Колекцията Fox-Cat* (от 91 детски разказа) и *Колекцията Dog-Birds* (от 142 детски разказа), в които са обхванати наративите на изследваните 121 деца по съответните картинни серии. При това данните са структурирани с оглед на отнесеността на изследваните лица към една от трите възрастови групи – деца във възраст съответно 3–4 години за първа група, 4–5 години за втора група, 5–6 години за трета група.

В основата на първата колекция от Корпус Б лежат две картинни истории *Cat Story* (Hickmann 2002) и *Fox Story* (Gülzow, Gagarina 2007), всяка от които се състои от по 6 черно-бели рисунки без текст – фигура 3 и 4.



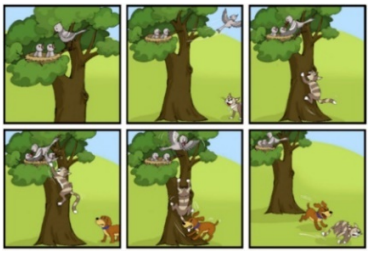
Фигура 3. Cat Story (Hickmann 2002)



Фигура 4. Fox Story (Gülzow, Gagarina 2007)

Втората наративна колекция на Корпус Б е базирана на две цветни картинни истории – *Baby Birds* и *Dog Story*, заимствани от MAIN: The Multi-

lingual Assessment Instrument for Narratives (Gagarina et al. 2012) – вж. фигура 5 и 6.



Фигура 5. Baby Birds Story (Gagarina et al. 2012)

Фигура 6. Dog Story (Gagarina et al. 2012)

Както става ясно от изложението, при създаването на корпус Б подборът на картинните истории и техният формат (черно-бял или цветен) не са базирани на случаен принцип. Използвани са специално разработени материали, които вече са използвани за създаването на еднотипни корпуси с наративи на много други езици, съотносими с унифицираната система MAIN, предназначена за анализ на детски наративи (Gagarina et al. 2012; Gagarina et al. 2019).

Данните и в двата подкорпуса на представяния тук корпус с българска детска реч са транскрибирани в унифицирания формат CHAT на системата CHILDES, което ги прави съотносими с корпусите на други езици от платформата. А пакетът от програми CLAN позволява да се направи различен тип анализ на въведените диалози и наративи. В този смисъл CLAN осигурява възможност автоматично да се получат най-разнообразни статистически и съдържателни резултати от транскрибираните и кодирани данни за честотата на думите, за лексикалното разнообразие и съчетаемост, за употребените в съответната речева сесия специфични думи и форми (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.

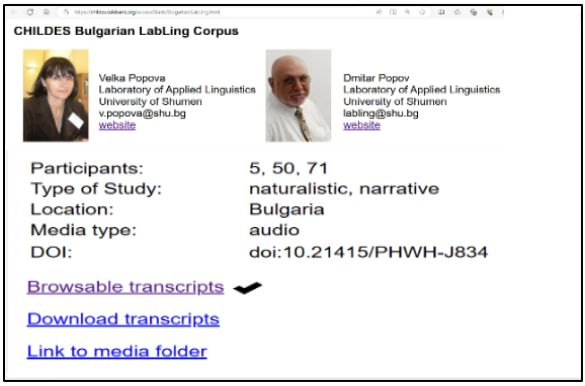
CHILDES е изключително пластична и удобна система, което предполага възможност нейните функционалности да бъдат достъпни за всеки изследовател. Дори и в случаите, когато все още корпусите не са публикувани на платформата и няма достъп до интернет. Достатъчно е на съответния компютър да са инсталирани програмите от пакета CLAN и да са налични необходимите речевите данни, транскрибирани в термините на CHILDES. Например много преди официално да се появи Bulgarian LabLing Corpus, системата беше използвана по този начин в проучвания на езиковото развитие на български деца, което пък на свой ред беше своеобразна апробация на приложимостта ѝ към нашия език. Днес обаче, когато българският корпус е официално публикуван, ученият на практика има свободата да избере дали да работи онлайн в самата система, като използва директно данните от рубриката Browsable transcripts, или да си свали на персоналното устройство данните рубриката Download transcripts.

За илюстрация на казаното на фигура 7 ще бъдат представени основните стъпките в алгоритъма на работа директно в платформата.

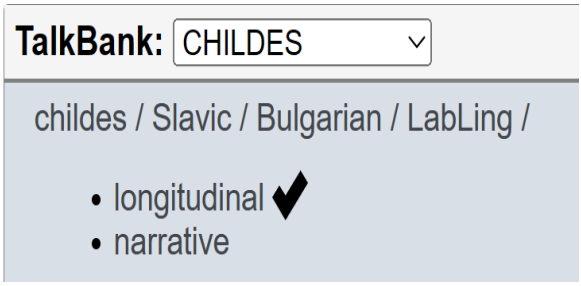
ПЪРВА СТЪПКА
Влизане в Bulgarian LabLing Corpus

https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html

ВТОРА СТЪПКА
Избор на Browsable transcripts



ТРЕТА СТЪПКА
Избор на корпус



ЧЕТВЪРТА СТЪПКА
Избор на изследвано лице

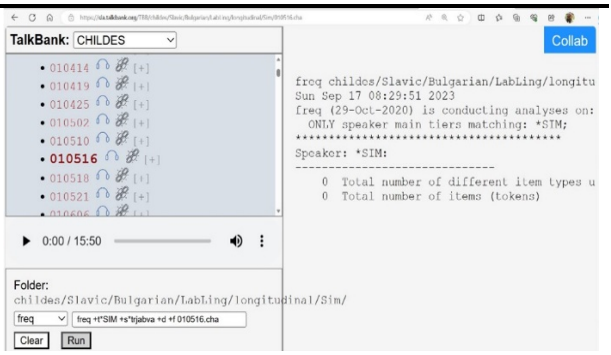
ПЕТА СТЪПКА
Избор на транскрипт

ШЕСТА СТЪПКА
И СЕДМА СТЪПКА

- **Задаване на формула (ДЕТЕ)**

freq +t*SIM +s*trjabva +d +f 010516.cha

- **Резултат**

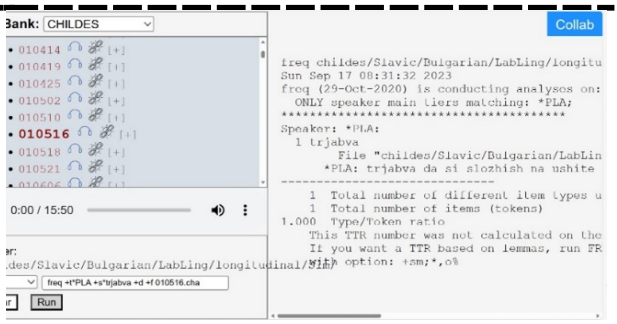


ОСМА И ДЕВЕТА СТЪПКА

- **Задаване на формула (ВЪЗРАСТЕН)**

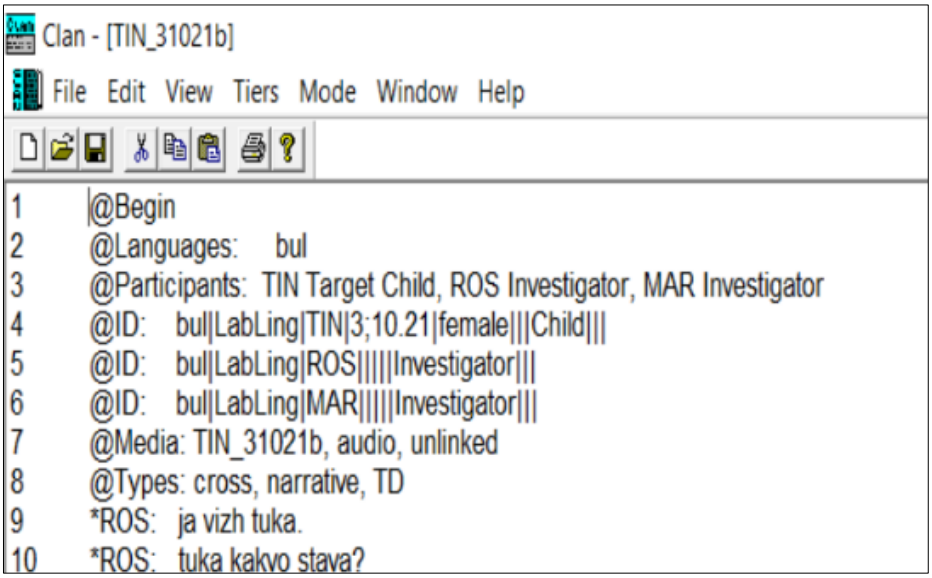
freq +t*PLA +s*trjabva +d + f 010516.cha

- **Резултат**



Фигура 7. Алгоритъм за работа с програмите CLAN директно в платформата CHILDES

От Фигура 7 се вижда, че да се работи директно в самата платформа, е не просто удобно, но и изключително лесно. Същевременно обаче системата CHILDES дава много възможности, една от които е да се изследват всички участници в общуването. Това ясно се откроява на фигура 7 при сравняването на шеста и седма стъпка с осма и девета стъпка, с които се извличат данни от един и същ транскрипт за фреквентността на модалния глагол *трябва* в речевата продукция на изследваното дете (SIM) и на неговата майка (PLA). Това е само едно свидетелство за широката приложимост на Bulgarian LabLing Corpus, която се определя от факта, че във всеки един от транскриптите са включени данни за идентификацията на изследваните лица (демографски и езикови параметри) и за типа на съответния корпус (лонгитудинален или кроссекционен). Вж. фигура 8 и фигура 9:



Фигура 8. Служебни редове на транскрипт в CHAT формат от кроссекционните данни на Корпус Б

При това не става дума просто за самоцелна идентификация на изследваните лица, а за възможност за извличане и анализ на речевите данни на всяко едно от тях или на подбрана група, което е изключително важно за онтогенетичните проучвания. Така се осигурява достъп и до *речта на възрастните към децата*, чрез която децата получават информация за системата на езика цел (т.нар. *позитивни входни данни*). А резултатите от анализа на подобни данни несъмнено осигуряват солидна база за изграждането на адекватни обяснителни модели на усвояването на езика, в които вече става възможно да се открои конкретният принос на възрастните и на тяхната реч в онтогенетичния процес при усвояването на определени езикови

феномени, както и да се изведат характеристиките на специфичния регистър, наричан *реч на възрастните към децата*.

Participants:

participant	role	name	language	age	sex
ALE	Target_Child	-	bul	2;02.24	-
STE	Sister	-	bul	8;01.20	-
SAN	Relative	Grandfather	bul	-	-
VEL	Mother	-	bul	-	-

```
@Begin
@Languages: bul
@Participants: ALE Target_Child, STE Sister, SAN Grandfather Relative, VEL Mother
@ID: bul|LabLing|ALE|2;02.24||||Target_Child|||
@ID: bul|LabLing|STE|8;01.20||||Sister|||
@ID: bul|LabLing|SAN|||||Relative|||
@ID: bul|LabLing|VEL|||||Mother|||
@Date: 22-APR-1991
@Tape Location: Cassete 10: Side 2 i komentirat
@Types: long, toyplay, TD
1    VEL: kakvo pravish sega ?
2    ALE: shija .
```

Фигура 9. Идентификационни данни на ИЛ в траскрипт от Bulgarian LabLing Corpus в системата CHILDES

Допълнителна изследователска перспектива тук може да се открие във връзка и с проучването на влиянието на останалите сублинги върху езиковото развитие на съответното дете. В това отношение метаданните за тях са от особена важност, което също е предвидено в системата CHILDES. Така сред участниците на интеракцията, представени на фигура 9, откриваме по-голямата сестра (STE) на изследваното дете, за която са посочени социална роля, пол, възраст, език.

В логиката на казаното дотук би могло да се обобщи, че публикуването на български данни с детска реч на платформата CHILDES е много важно и полезно за изследователите, тъй като им осигурява възможност за проучвания на българския език в контекста на една оптимална работна среда, характеризираща се със свободен достъп до множество разнообразни данни, както и пакет с компютърни програми, чрез който огромният емпиричен масив може да бъде обработван статистически.

Същевременно CHILDES предлага един от оптималните електронни формати на съвременната корпусната лингвистика, даващ възможност за мултимодално интегрирано многоаспектно представяне на речевото поведение, а това вече е гаранция за получаването на обективни резултати, защото осигурява възможност изследваният феномен да бъде наблюдаван едновременно от различни гледни точки. То е особено необходимо при изучаването на феномена *говорещо дете* поради специфичните възрастови

особености и ограничената лингвистична и комуникативна компетенция в зората на човешкия живот.

Във връзка с това е важно да се отбележи приносът на Брайън МакУини (вж. MacWhinney 2007; MacWhinney & Wagner 2010). Той именно разработва интерактивната мултимодална система на платформата CHILDES специално за целите на анализа на речта и оптималното ѝ многоаспектно визуализиране, в резултат на което се постига ново качество в проучванията на компонентите на невербалната комуникация и влиянието на прагматичния контекст в речта на детето. Унифицираното анотиране на екстралингвистичните данни, които съпътстват речта на наблюдаваните деца, както и непрекъснатият режим на връзка между транскриптите и съответните аудиофайлове създават възможности и перспективи не само за изолирано изследване на езиковите феномени, но и за пълноценното им цялостно проучване в динамичния режим на речевото общуване. С помощта на CHILDES речевото поведение на изследваните деца може да бъде представено в мултимодална перспектива.

Тази иновативната система на Б. МакУини за интегрираното представяне на речевите данни е приложима към представяния български корпус с детска реч, доколкото в голямата си част конкретните транскрипти са съпътствани от аудиофайлове. Така на един следващ етап те биха могли да бъдат синхронизирани, при което ще се визуализират изпълнимият аудиофайл и неговата чувана в реално време звукова реализация в SONIC MODE, изобразена като осцилограма в долния край на прозореца на програмата CLAN, както е демонстрирано в извадката на фигура 10:

Фигура 10. Аудиотранскрипция в SONIC MODE

Възможността за многократно аудиореконструиране на кой да е фрагмент от транскриптите от корпуса дава гаранция за пълноценен и адекватен анализ в процеса на проучване както на такива комплексни езикови феномени и категории, като например модалността, за интерпретацията на които е важна и интонацията, така също и на фонетичната страна на детската реч. В това отношение трябва да се отбележи още едно предимство на CHILDES, а именно операционната съвместимост на нейния пакет програмни ресурси CLAN с други системи за обработка на реч, като например: Praat, EXMARa LDA, PHON и др.

Заключение

В настоящата работа беше направен опит да се открият ползите на интерактивната система CHILDES. Общопризнато е, че тя осигурява условия за по-голяма прецизност при събирането, транскрибирането и кодирането на данните, а също така предлага инструменти за автоматизиран анализ на големи количества разговорен материал, което разширява емпиричната база, върху която се строят новите теории. Към това в логиката на предложеното тук изложение трябва да се добави още един акцент, а именно, 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 за изследване на всички аспекти на речевата интеракция.

С публикуването на българските данни на платформата CHILDES се постига разширяване на възможностите ѝ за кослингвистичните изследвания с още един славянски език. А българската лингвистична традиция се обогатява с още един универсален комфортен стандарт за изследване на езиковата онтогенеза, благодарение на който биха могли бързо, точно и надеждно да се осъществяват съпоставки с голямо количество езици. На тази база могат да се строят солидни типологии и модерни теории.

Разбира се, *Bulgarian LabLing Corpus* е само отделен миниатюрен фрагмент от една многоезична виртуална мозайка, която непрекъснато се разширява и обогатява, тъй като тя се изгражда в термините 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 като положителен знак може да се посочи интеграцията на американския CMU-TalkBank в европейския CLARIN (https:// talkbank.org/knowledge/), част от която е и коментираният тук първи български CHILDES корпус, създаден в рамките на работния план на КлаДа-БГ.

Благодарности

Това изследване е частично финансирано от Българската национална интердисциплинарна изследователска инфраструктура за ресурси и технологии в полза на българския език и българското културно наследство, част от европейските инфраструктури CLARIN и DARIAH – КЛаДА-БГ, договор Д01-167/28.07.2022.

ЛИТЕРАТУРА

Кёстер-Тома 1993: *Кёстер-Тома, З.* Стандарт, субстандарт, нонстандарт. – *Русистика*. Берлин, № 2, с. 15–31.

Младенов 1934: *Младенов, Ст.* Няколко езикословни въпроси у проф. Ив. А. Георгов в работите му за развоя на детския говор. – *Годишник на СУ. ИФФ*, кн. XXX.

Попов 2005: *Попов, Д.* Аспекти на българското произношение в сферите на стандарта, субстандарта и нонстандарта. – *Стандарт и субстандарт – диахронни и синхронни аспекти*. Шумен, УИ „Епископ Константин Преславски“, с. 191–204.

Попов 2016: *Попов, Д.* Лингвистична персонология. Шумен, УИ „Епископ Константин Преславски“.

Попов, Попова 2021: *Попов, Д., В. Попова.* Феноменът детска реч в светлината на многостепенния модел на българската реч. – *Отговорността пред езика*. Книга 7. Шумен, УИ „Епископ Константин Преславски“, сл. 258–275.

Стоянова 2022: *Стоянова, Ю.* Проблеми на психолингвистиката. София, УИ „Св. Климент Охридски“.

Brown, 1973: *Brown, R*. A first language: The early stages. Cambridge, MA: Harvard University Press, pp. 437.

Gagarina et al. 2012: *Gagarina, N., D. Klop, S. Kunnari, K. Tantele, T. Välimaa, I. Balciuniene, U. Bohnacker and J. Walters.* Multilingual Assessment Instrument for Narratives (MAIN). ZAS papers in linguistics, 56, pp.1–140.

Gagarina et al. 2019: *Gagarina, N., D. Klop, S. Kunnari, K. Tantele, T. Välimaa, I. Balciuniene, U. Bohnacker and J. Walters.* MAIN: Multilingual Assessment Instrument for Narratives – Revised. – *ZAS Papers in Lin*guistics, 63, p. 20.

Gülzow, Gagarina 2007: *Gülzow, I., N.Gagarina.* Intersentential pronominal reference in child and adult language. – In: *ZAS Papers in Linguistics*, Nr. 48, pp. 203−223.

Hickmann 2002: *Hickmann, M*. Children's Discourse: Person, Space and Time across Languages. – *Cambridge Studies in Linguistics* (Vol. 98). Cambridge University Press, p. 412.

MacWhinney 2000: *MacWhinney*, *B*. The CHILDES Project. Tools for Analyzing Talk. Third Edition. Mahwah, NJ, Lawrence Erlbaum Ass., p. 366.

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

MacWhinney & Wagner 2010: *MacWhinney, B. & J. Wagner*. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. – In: Gesprächsforschung – *Online-Zeitschrift zur verbalen Interaktion*, Ausgabe 11. 2010, pp. 154–173.

REFERENCES

Brown, 1973: *Brown, R*. A first language: The early stages. Cambridge, MA: Harvard University Press, 1973. pp. 437.

Gagarina et al. 2012: *Gagarina, N., D. Klop, S. Kunnari, K. Tantele, T. Välimaa, I. Balciuniene, U. Bohnacker and J. Walters.* Multilingual Assessment Instrument for Narratives (MAIN). ZAS papers in linguistics, 56, pp.1–140.

Gagarina et al. 2019: *Gagarina, N., D. Klop, S. Kunnari, K. Tantele, T. Välimaa, I. Balciuniene, U. Bohnacker and J. Walters.* MAIN: Multilingual Assessment Instrument for Narratives – Revised. – *ZAS Papers in Lin*guistics, 63, p. 20.

Gülzow, Gagarina 2007: *Gülzow, I., N.Gagarina.* Intersentential pronominal reference in child and adult language. – In: *ZAS Papers in Linguistics*, Nr. 48, pp. 203−223.

Hickmann 2002: *Hickmann, M*. Children's Discourse: Person, Space and Time across Languages. – *Cambridge Studies in Linguistics* (Vol. 98). Cambridge University Press, p. 412.

Kjoster-Toma 1993: *Kjoster-Toma, Z.* Standart, substandart, nonstandard. – *Rusistika*, N 2, s. 15–31.

MacWhinney 2000: *MacWhinney*, *B*. The CHILDES Project. Tools for Analyzing Talk. Third Edition. Mahwah, NJ, Lawrence Erlbaum Ass., p. 366.

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

MacWhinney & Wagner 2010: *MacWhinney, B. & J. Wagner*. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. – In: Gesprächsforschung – *Online-Zeitschrift zur verbalen Interaktion*, Ausgabe 11. 2010, pp. 154–173.

Mladenov 1934: *Mladenov, St*. Nyakolko ezikoslovni vaprosi u prof. Iv. A. Georgov v rabotite mu za razvoya na detskiya govor. – *Godishnik na SU. IFF*, kn. XXX.

Popov 2005: *Popov, D.* Aspekti na balgarskoto proiznoshenie v sferite na standarta, substandarta i nonstandarta. – *Standart i substandart – diahronni i sinhronni aspekti*. Shumen, UI "Episkop Konstantin Preslavski", s. 191–204.

Popov 2016: *Popov, D*. Lingvistichna personologiya. Shumen, UI "Episkop Konstantin Preslavski".

Popov, Popova 2021: *Popov, D., V. Popova*. Fenomenat detska rech v svetlinata na mnogostepenniya model na balgarskata rech. – *Otgovornostta pred ezika*. Kniga 7. Shumen, UI "Episkop Konstantin Preslavski", s. 258–275.

Stoyanova 2022: *Stoyanova, Yu*. Problemi na psiholingvistikata. Sofia, UI "Sv. Kliment Ohridski".

✉ *Проф. д-р Велка Попова*, Катедра по български език
Шуменски университет „Епископ Константин Преславски"
ул. „Университетска" № 115А, 9700 Шумен, България

✉ *Prof. Velka Popova, PhD*, Department of Bulgarian Language
Konstantin Preslavski University of Shumen
115A Universitetska St., Shumen 9700, Bulgaria