УДК 8123
Попова Велка Александрова
Доктор филологии, профессор
Профессор, Шуменский университет имени Константина Преславского, Кафедра болгарского языка, Лаборатория прикладной лингвистики
e-mail: v.popova@shu.bg

Попов Димитър Димитров
Доктор филологии, профессор
Профессор, Шуменский университет имени Константина Преславского, Кафедра болгарского языка, Лаборатория прикладной лингвистики
e-mail: labling@shu.bg

БЪЛГАРСКИ КОРПУС С ДЕТСКА РЕЧ НА ПЛАТФОРМАТА *CHILDES*

Аннотация. Статья посвящена первому болгарскому компьютерному корпусу детской речи – Bulgarian LabLing Corpus, который был опубликован на платформе CHILDES.
Ключевые слова. языковой корпус, детский язык, CHILDES, болгарский

Popova Velka Aleksandrova
Ph.D., professor
Konstantin Preslavsky University of Shumen, Faculty of Humanities, Department of Bulgarian Language, Laboratory of Applied Linguistics, Professor
e-mail: v.popova@shu.bg

Popov Dimitar Dimitrov
Ph.D., professor
Konstantin Preslavsky University of Shumen, Faculty of Humanities, Department of Bulgarian Language, Laboratory of Applied Linguistics, Professor
e-mail: labling@shu.bg

BULGARIAN CHILD LANGUAGE CORPUS ON THE CHILDES PLATFORM

Abstract. The article is dedicated to the first Bulgarian computer-based corpus of children`s speech – the Bulgarian LabLing Corpus, which has been published to the CHILDES platform
Keywords. language corpus, child language, CHILDES, Bulgarian

Увод

Детската реч е източник на важна информация за механизмите на усвояване и използване на езика в процеса на речевото общуване. В нейното изследователско поле се проверяват фундаментални хипотези за вродената и придобитото в езика, за ролята и мястото на езика в познавателното развитие на човека и др. Всичко това предполага необходимостта от солидна емпирична операционализация на създадените теоретични конструкти, а това означава събирането на огромно количество речеви продукти и тяхното изследване в светлината на надеждни и адекватни научни подходи.

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 лингвистичното обяснение на езиковите правила и принципи към търсенията 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 в научното пространство 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" [6: 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, а именно – платформите 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), създаден в Лабораторията по приложна лингвистика към Шуменския университет. 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 предположения 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. 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 систематичните 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. И до днес тези своеобразни „бебешки биографии“ не губят своята значимост. През 30-те години на XX в. в рамките на бихевиоризма се реализират първите срезови проучвания, в които вече могат да се сравняват образци от много деца в 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 лонгитудни те срезови изследвания, която е своеобразен синтез на методологичните достижения 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 преодолее фрагментарността и случайността, присъща за дневниците и на срезовите данни.

В зависимост от поставената цел отделните методи имат своите предимства и недостатъци. Така 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. Лонгитудни те 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 лонгитудни 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 технически средства и програмни продукти. 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 крослингвистични 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).

В обобщение на казаното в този параграф може да се направи изводът, че създаването на CHILDES бележи апогеят 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. 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 да бъдат и 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 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 в свободна за безплатен достъп зона в интернет.

CHILDES представлява система за обмен на данни по 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 [1] 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.

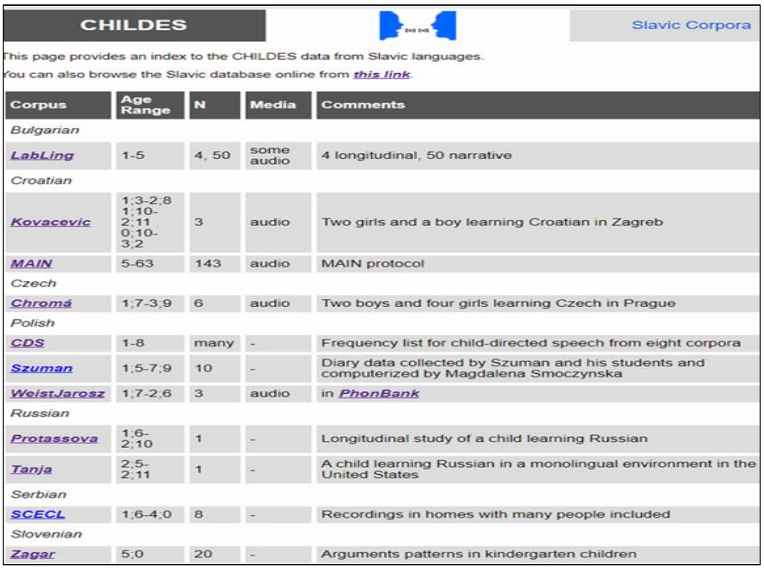
Базата от данни на CHILDES съдържа голям обем от сведения за усвояване на множество езици, като в нея има и специален раздел за аномалиите в езиковото развитие и за усвояването на втори език.

Несъмнена е ползата от автоматизираната компютъризирана 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 количества разговорен материал, което разширява значимо 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.

3. *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* (вж. Фиг. 1).



CHILDES

Slavic Corpora

This page provides an index to the CHILDES data from Slavic languages.
You can also browse the Slavic database online from this link.

Corpus	Age Range	N	Media	Comments
Bulgarian				
LabLing	1-5	4, 50	some audio	4 longitudinal, 50 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;7-3;9	6	audio	Two boys and four girls learning Czech in Prague
Polish				
CDS	1-8	many	-	Frequency list for child-directed speech from eight corpora
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
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	-	Recordings in homes with many people included
Slovenian				
Zagar	5;0	20	-	Arguments patterns in kindergarten children

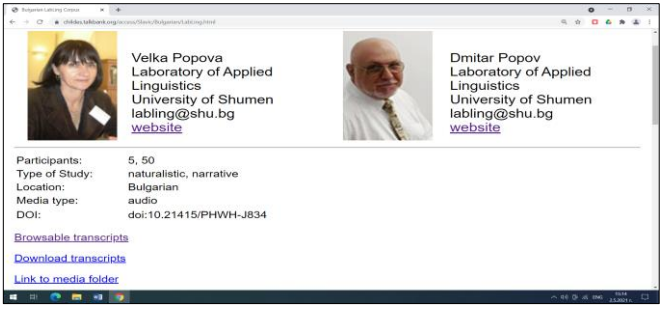
Фиг. 1. Славянски корпуси, публикувани на платформата CHILDES

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

Представяният тук корпус с българска детска реч е резултат от дългогодишния труд на изследователи от LABLING. Данните на корпуса са

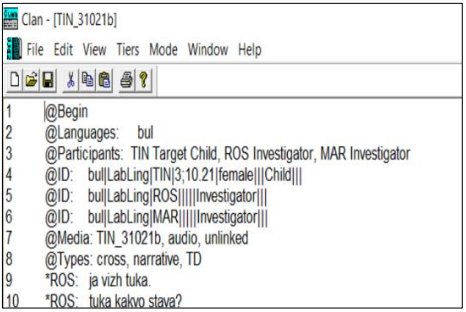
транскрибирани в унифицирания формат CHAT на системата CHILDES [5], което ги прави съотносими с корпусите на други езици от платформата. А пакетът от програми CLAN позволява да се направи различен тип анализ на въведените диалози (фонетичен, морфологичен, синтактичен) и коментарите към тях. В този смисъл CLAN осигурява възможност автоматично да се получат най-разнообразни статистически и съдържателни резултати от транскрибираните и кодирани данни, като например за честотата на думите, за лексикалното разнообразие и съчетаемост, за употребените в съответната речева сесия специфични думи и форми (като 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.

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



Фиг. 2. Българският корпус с детска реч на CHILDES

Широката приложимост на Bulgarian LabLing Corpus е предпоставена 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. Фиг. 3:



Фиг. 3. Служебни редове на един транскрипт в CHAT-формат

В структурата на Bulgarian LabLing Corpus са обособени два подкорпуса, единият от които е лонгитудинален, а другият – кроссекционен:

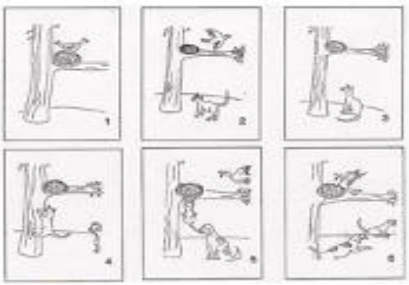
- Корпус А с лонгитудни те данни на 5 български момичета в ранна възраст;
- Корпус Б с наративи на 50 деца в предучилищна възраст, който може да се определи като кроссекционен, доколкото изследваните лица допълнително са разделени по възраст в три групи.

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;04.14–2;00.09). 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). Аудиозаписите на две от децата (ALE и TEF) са направени от изследователския екип на LABLING, а на другите три (ELI, BOG и SIM) – от техните майки. Транскрипцията и кодирането на материала са осъществени от изследователския екип на LABLING.

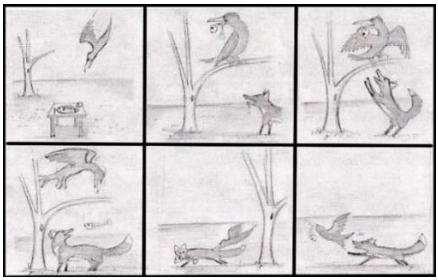
Корпус Б съдържа 91 транскрипта на детски наративи, извлечени от 50 деца монолингви (носители на български език). Те са записвани на диктофон в няколко детски градини в Шумен и Варна (Североизточна България), като само в няколко единични случая – вкъщи или на улицата. В този кроссекционен корпус децата са разпределени в 3 групи по възраст:

- първата група обхваща 21 деца във възраст 3–4 години (От включените в корпуса **36** наратива 21 са представени само с транскрипти, а останалите 15 – както с транскрипти, така и с аудиофайлове.)
- втората група обхваща 23 деца във възраст 4–5 години (От включените в корпуса **43** наратива 10 са представени само с транскрипти, а останалите 33 – както с транскрипти, така и с аудиофайлове.);
- третата група обхваща 6 деца във възраст 5–6 години (включените в корпуса **12** наратива са представени и с транскрипти, и с аудиофайлове).

В основата на Корпус Б лежат две картинни истории, всяка от които се състои от по 6 черно-бели рисунки без текст. Това са Cat Story и Fox Story (разработени от изследователския екип на ZAS-Берлин, ръководен от Дагмар Битнер, и публикувани за първи път от I. Gülzow и N. Gagarina [3] – Фиг. 4 и Фиг. 5).

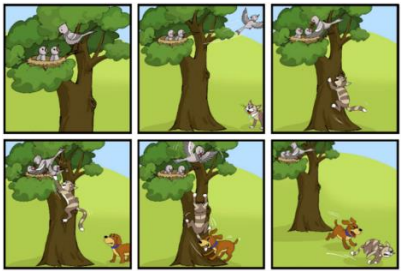


Фиг. 4. Cat Story [4]

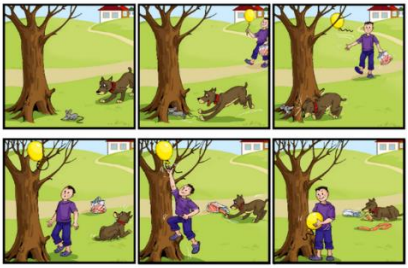


Фиг. 5. Fox Story [2]

Bulgarian LabLing Corpus е в процес на разширяване с нови данни. Изследователският екип на LABLING вече работи по събирането и транскрибирането на нови лонгитудни данни, с които да се обогати Корпус А. Колекцията с наративи (Корпус Б) се обогатява с разкази по две нови цветни картинни истории – Baby Birds and Dog Story, заимствани от MAIN: The Multilingual Assessment Instrument for Narratives [2] – Вж. Фиг. 6 и Фиг. 7. На този етап са направени звукови регистрации на 80 наратива на 40 деца в предучилищна възраст, които все още не са транскрибирани, тъй като продължава процесът на акумулиране на нови данни.



Фиг. 6. Baby Birds Story



Фиг. 7. Dog Story

Подборът на картинните истории и техният формат (черно-бял или цветен) не е случаен. Както стана ясно, това са специално разработени материали, които не само са използвани за създаването на корпуси с наративи на много езици, но и в някои случаи са свързани с унифицираната система *MAIN,* предназначена за анализ на детски наративи [2].

Публикуването на българските 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 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.

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 ДО1-272/16.12.2019

Заключение

Bulgarian LabLing Corpus е само един 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. Същевременно вече е факт интеграцията на CMU-TalkBank в CLARIN (https://talkbank.org/knowledge/).

Литература

1. Brown, R. A first language: The early stages. Cambridge, MA: Harvard University Press, 1973. p. 437.
2. Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balciuniene, I., Bohnacker, U. and Walters, J. Multilingual Assessment Instrument for Narratives (MAIN). ZAS papers in linguistics, 2012, 56, pp.1-140. / Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., Walters, J. MAIN: Multilingual Assessment Instrument for Narratives–Revised. ZAS Papers in Linguistics, 2019, 63, p. 20.
3. Gülzow, I., Gagarina, N. Intersentential pronominal reference in child and adult language. In: ZAS Papers in Linguistics, Nr. 48, 2007, pp. 203−223.
4. Hickmann, M. Children's Discourse: Person, Space and Time across Languages. Cambridge Studies in Linguistics (Vol. 98). Cambridge University Press. 2002, p. 412.
5. MacWhinney, B. The CHILDES Project. Tools for Analyzing Talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Ass. 2000, p. 366.
6. Попов, Д. Лингвистична персонология. Шумен: Университетско издателство „Епископ Константин Преславски". 300 с.