



Latvijas Universitātes
Matemātikas un informātikas institūts



Meklēšana ortogrāfiski un morfoloģiski marķētā bērnu runas korpusā

*Ilze Auziņa, Kristīne Levāne-Petrova,
Baiba Saulīte, Artūrs Znotiņš, Roberts Darģis*

LU Matemātikas un informātikas institūts



Projekts ir saņēmis finansējumu no Norvēģijas finanšu instrumenta 2009. -2014.
gadam saskaņā ar projekta līgumu nr.NFI/R/2014/053

LAMBA

Bērnu runas korpuss

Projekts “Latviešu valodas monolingvāla un bilingvāla apguve: rīki, teorijas un lietojums”

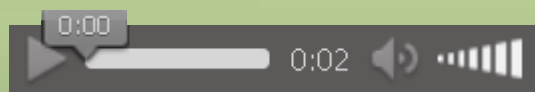
- Projekta ilgums: 03.2015.–04.2017.
- 3 monolingvāli latviski runājošu bērnu runas apakškorpusi
- 1 bilingvāli – latviski un krieviski – runājoša bērna runas apakškorpuss
- Datu atšifrēšana (ortogrāfiskā transkribēšana) un morfoloģiskā marķēšana

Datu iegūšana

- Bērnu runa tika ierakstīta ik nedēļu 17 mēnešus
- Vidējais katra ieraksta ilgums:
 - Monolingvāli bērni – 30 minūtes
 - Bilingvāls bērns – 30 minūtes katrā valodā
- Sasniegtais apjoms: aptuveni 134 stundas
- Bērnu vecums: 17–45 mēneši
- Ieraksta ierīces: diktofons vai mobilais telefons
- Ieraksta vide: bērnu dzīvesvieta
- Par katru ierakstu tiek saglabāta papildu informācija, piem., ieraksta vieta, nodarbe

Korpusa izveides posmi

Audio / video



Ortogrāfiskā transkripcija

valdītes [vardītes] dzied {xxx}

Morfoloģiski marķēti dati

vardītes//valdītes

ncfpn5

dzied

vmnipt330an

{xxx}

xx

Galvenie ortogrāfijas transkribēšanas principi bērnu runas korpusā (1)

Galvenās marķējuma kodu grupas, aprakstot bērnu runu, ir šādas:

- speciāli valodas apguvēju formu marķieri, piemēram, speciāli apzīmējumi bērnu jaundarinājumiem,
- neskaidru / nesaprotamu valodas vienību kodi,
- nepabeigtu vārdu kodi,
- šablonisks vārdu lietojums,
- standartizētas pareizrakstības normas.

Galvenie ortogrāfijas transkribēšanas principi bērnu runas korpusā (2)

- Atveidojot mutvārdu valodas piemērus rakstu formā, saglabāts runas fakta autentiskums, piemēram,
 - **jaunvārdi, bērna radītās vārdformas un skaņu kopas** tiek norādītas, pirms tiem rakstot simbolu & un kvadrātiekvāš dodot atbilstošo literārās valodas ekvivalentu: &*vauva* [suns], &*jajā* [zaķis], &*mā* [māja];
 - **neskaidrs teksts** tiek likts figūriekavās, norādot iespējamo vārdu vai vārda daļu: {*teksts*}, {*xxx*};
 - **bērna čalošanas** (*babbling*) apzīmēšanai tiek izmantots apzīmējums .
- Lielie sākumburti tiek izmantoti tikai īpašvārdu pierakstīšanai.
- Runas piemēri tiek segmentēti, un tiek izmantotas atbilstošās interpunkcijas zīmes.
- Transkribējot bilingvāla bērna runu, tiek norādīta pārslēgšanās no vienas valodas uz citu.

Galvenie ortogrāfijas transkribēšanas principi bērnu runas korpusā (3)

- **Smieklus** apzīmēšanai tiek izmantots simbols @. Apzīmējums tiek atkārtots tik reižu, cik zilbēm atbilst smiešanās. Ja teksts tiek izrunāts smeļoties, tiek norādīts smeļoties izrunātā teksta sākums un beigas: <@> *teksts* </@>.
- **Fizioloģisks troksnis** — cilvēka radīts troksnis (šņaukšanās, mēles klakšķināšana, čāpstināšana, žagošanās, šķavas u. tml.) tiek norādīts, izmantojot birku <ftr>, <ftr> *teksts* </ftr>.
- **Troksnis**, ko rada fonā runājoši cilvēki vai darbojošās ierīces, piemēram, radio, televizors, braucoša automašīna; arī neidentificētas skaņas ierakstā, tiek norādīts, izmantojot birku <tr>.
- **Pauzes** – gan aizpildītas, gan neaizpildītas – tiek norādītas, apaļajās iekavās liekot punktu (.) vai norādot pauzes ilgumu sekundēs (0.54).

Morfoloģiskā marķēšana

- Pēdējais korpusa izveides posms.
- Izmantots jau iepriekš latviešu valodai izstrādātais morfoloģiskās anotēšanas standarts un jau esošie morfoloģiskās marķēšanas rīki.
- Izmantota gandrīz visa esošā morfoloģisko apzīmējumu kopa (MULTEXT-EAST), izņemot dažas semantiskas adverbu pazīmes.
- Marķēšana ir veikta automātiski.
- Esošais morfoloģiskais marķējums ir transformējams MOR gramatikā – programmā, kas nodrošina korpusu automātisku marķēšanas metodi CHILD formātā.

Morfoloģiskā marķējuma piemērs

kājas

ncfpn4

n – lietvārds

c – sugas vārds

f – sieviešu dzimte

p – daudzskaitlis

a – nominatīvs

4 – 4. deklinācija

sapinās

vmyisi130an

v – darbības vārds

m – patstāvīgs

y – atgriezenisks

i – īstenības izteiksme

s – pagātne

i – nepārejošs

1 – konjugācija

3 – persona

0 – nepiemīt

a – darāmā kārtā

n – nav noliegtais darbības vārds

Konkrētā meklēšana

- Tiek meklēta konkrēta vārdforma.

The screenshot shows a search interface with a search bar containing the word "tētis". Below the search bar are buttons for "Meklēt", "Eksportēt", "Notīrīt", "Slēpt anotācijas", and "Slēpt tagus". To the right, there are filters for "Runātājs" (bērns, tēvs, māte, māsa, cits), "Dzimums" (vīr., siev., -), and "Vecums" (-, -, -). The results section shows "Tika atrasti 11 rezultāti (0.08s)". The results table has columns for information icons, text, and actions. The results are:

Info	Text	Action
i	tētis	
i	tētis	
i	tētis	apgāzās
i	tētis	apgāzās
i	ui (.)	tētis
		atradās/atladās

- Tiek meklēts noteikts teikuma komunikatīvais tips, konkrēti – jautājuma teikums vai izsaukuma / pamudinājuma / vēlējuma teikums.

The screenshot shows a search interface with a search bar containing the word "?". Below the search bar are buttons for "Meklēt", "Eksportēt", "Notīrīt", "Slēpt anotācijas", and "Slēpt tagus". To the right, there are filters for "Runātājs" (bērns, tēvs, māte, māsa, cits), "Dzimums" (vīr., siev., -), and "Vecums" (-, -, -). The results section shows "Tika atrasti 3062 rezultāti (0.52s)". The results table has columns for information icons, text, and actions. The results are:

Info	Text	Action
i	ar lauku omīti	?
i	kā viņai gāja	?
i	kā viņai gāja ? viņa mācēja salikt	?
i	tu viņai palīdzēji drusciņ	?
i	ja	?
i	šeit ir augša vai šeit	?

Aptuvenā meklēšana

Vaicājumu sintakse

Simboli	Raksturojums	Piemēri
%	neviens, viens vai vairāki simboli	<i>mamm%</i> → <i>mamma, mammu, mammamm</i> u. tml. <i>%a</i> → <i>a, ņa, ķa, jāja, mamma</i> u.tml. <i>ābol%</i> → <i>ābols, ābolu, āboliņš, ābolītis</i> u.tml
_	viens simbols	<i>ābol_</i> → <i>ābols, ābolu, ābolā</i> u. tml. <i>ābol__</i> → <i>ābolus, ābolam</i> u. tml.
%//%	tiek meklēta neprecīza izruna, kļūdaina forma (vispirms tiek norādīta vēlamā forma, aiz slīpsvītrām – bērna runā lietotā)	<i>%//%</i> → <i>trokšņa//tokšņa, tur//tu, ir//il</i> u.tml. <i>zaķis//%</i> → tiek atrasti visi varianti, kā bērnu runā tiek saukts <i>zaķis</i>

Meklēšana, izmantojot metadatus

Meklēšanā iespējams izmantot informāciju par

- runātāja dzimumu,
- vecumu (tikai bērnu vecums),
- sociālo lomu ģimenē (bērns, tēvs, māte u. tml.).

Runātājs	☒ bērns	☐ tēvs	☐ māte	☐ māsa	☐ cits
Dzimums	☐ vīr.	☐ siev.	☒ -		
Vecums	3 g. - 4 g. 1 mē				

Meklēšana morfoloģiski marķētā korpusā

- Iespējams atrast visus vienas vārdšķiras vārdus

Vaicājums	Raksturojums	Komentāri
T:n%	Tiek meklēti lietvārdi	Tiek atrastas arī skaņu kopas, bērnu jaundarinājumi, kas tiek lietoti konkrētās vārdšķiras vārda funkcijā, piem., <i>ģēbt</i> 'ģērbt', <i>kīt</i> 'krīt', <i>kausī</i> 'klausīties'
T:v%	Tiek meklēti darbības vārdi	
T:s%	Tiek meklēti saikļi	
		Tā kā latv. val. ir morfoloģiski daudznozīmīga, automātiskas marķēšanas laikā nav iespējams pilnīgi pareizi noteikt konkrētas vārdformas morfoloģisko raksturojumu

- Var meklēt viena vārda visas formas

Vaicājums	Raksturojums	Komentāri
L:braukt	Tiek meklētas visas darbības vārda <i>braukt</i> formas	Piedāvā arī izrunas variantus, kas sastopami bērna runā, piem., tiek atrasts arī izrunas variants <i>blauc</i> 'brauc', <i>bļaukt</i> 'braukt'
L:mamma	Meklē visas lietvārda <i>mamma</i> formas	

Bērnu runas korpuss

Secinājumi

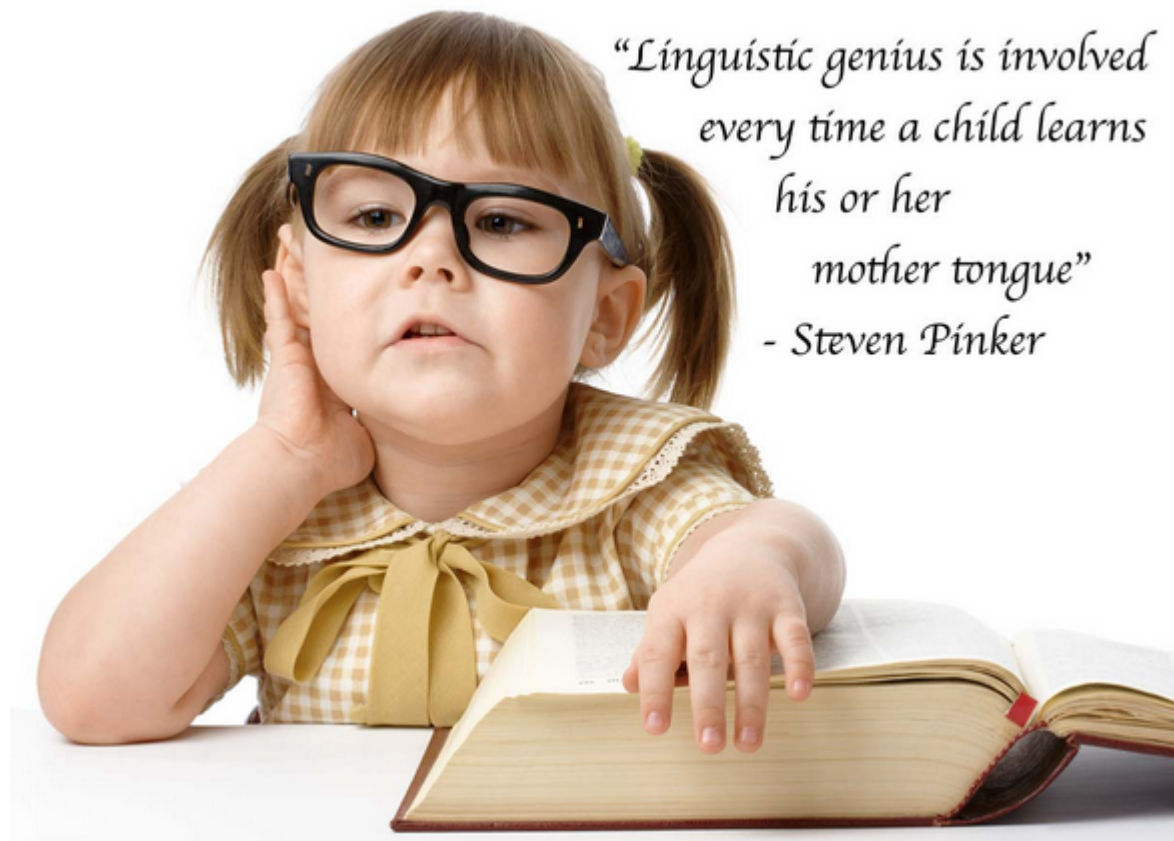
Iespēja meklēt bērnu runas korpusā ir pamats tālākiem pētījumiem un iegūto datu kvalitatīvai un kvantitatīvai analīzei.

Morfoloģiski marķētā korpusā, izmantojot atbilstošus korpusa programmrīkus, analīzei nepieciešamos datus var atrast ātrāk un precīzāk nekā vienkāršā elektronisko tekstu masīvā.

Izmantojot morfoloģiski marķētu korpusu, ir iespējams pētīt,

- kā bērns / kādā secībā apgūst gramatiskās kategorijas;
- kā attīstās bērnu vārdu krājums, kāds tas ir konkrētā vecumposmā;
- kāda skaņas vai skaņu kopas bērnu runā noteiktā vecumposmā tiek aizstātas ar citām skaņām.

Projekta mājaslapa: <http://lamba.lv>



*“Linguistic genius is involved
every time a child learns
his or her
mother tongue”
- Steven Pinker*