

Gyermeknyelvi korpuszok

Beszéd és Nyelvelemző Szoftverek
Szeged 2018

Babarczy Anna



- CHILES nemzetközi spontánbeszéd és történetmesélés korpusz (Brian MacWhinney, Carnegie Mellon)
- MONYEK magyar irányított interjú és történetmesélés korpusz (Mátyus Kinga, MTA Nyelvtudományi Intézet)
- GABI gyereknyelvi beszédadatbázis és információtár (Bóna Judit, ELTE)

CHILDES

- Child Language Data Exchange System, 1984!
- Újabban a TalkBank projekt része (nyelvtanulás és felnőtt beszélgetés)
- Részei: adatbázis + elemző eszközök
- Szabad hozzáférésű - hivatkozni illik az adatgyűjtő megadott cikkére
- Bárki beküldheti a saját adatait feltéve hogy a CHILDES formátumát követi

Adatbázis

- Etikai megfontolások: Az átiratok aninimizálva vannak - tulajdonnevek törölve
- Online, letölthető
 - A letölthető adatbázisban található a hivatkozandó cikkek
- Sajátos kategorizálás:
Celtic, Chinese, Clinical ... French, Frogs, German...
- Magyar anyag az Other kategóriában



A Frog Story

Frog, where are you (Mercer Mayer, 1969)

Átirat + hang/videó

- Beszéd átirata a CHILDES saját formátumában: CHAT
 - [Néhol az átirathoz audio vagy video van linkelve](#)
 - [Vagy nincs belinkelve, és a több részből álló hanganyaghoz van egy dummy file](#)

CHAT

@Loc: Other/Hungarian/Reger/021126.cha

1 @PID: 11312/c-00027779-1

2 @Begin

3 @Languages: hun

4 @Participants: CHI Target_Child , MOM Mother , SIS Orsi Sister

5 @ID: hun|Reger|CHI|2;11.26|||Target_Child|||

6 @ID: hun|Reger|MOM||||Mother|||

7 @ID: hun|Reger|SIS||||Sister|||

8 @Media: 021126, audio, unlinked

9 @Date: 12-SEP-1993

10 @Tape Location: Tape XXI , Side B. 000.

11 @Transcriber: Szilvia Papp.

12 @Situation: Mother and Miki talking while a baby is crying in the background.

13 @Comment: No original transcription for this.

14 *MOM: <dehúznak> [?] Miki .

15 *CHI: hát jó piac .

16 *MOM: aztán ?

17 *CHI: aztán bú [//] vagy hadd dugjak már ide [//] csigabigát .

18 *MOM: igen , aztán ?

19 *CHI: kagylót .

20 *MOM: kagylot is ad ?

21 *CHI: igen (.) előtt !

697 @End

CHAT tierek

- Fő tier: *CHI: Gyerek mondata, egy mondat egy sor.
- Altier: %nev: A fő tierhez tartozó kódok
 - A %mor tier: automatikus morfológiai elemzés a CHILDES morfológiai rendszere szerint
 - A %gra tier: automatikus grammatikai elemzés a CHILES nyelvtana szerint - spanyol példa
- Ezek sajnos a magyarra nem használhatók
- De bármilyen tier-t ki lehet találni magáncélokra

CLAN

- Átírást segítő “szövegszerkesztő”
 - offline működés
 - letölthető Windows, Mac, Unix verzió
- Egy sor elemzőprogram

Apple Safari File Edit View History Bookmarks Window Help Thu 6:58 pm

Commands

working /Applications/CLAN/work/
output
lib /Applications/CLAN/lib/
mor lib /Applications/CLAN/lib/
Progs ?
Recall 18oct18 Run

/Users/anna/Desktop/ChildCorpora/hf_5_g.cha

```
1 @Begin
2 @Languages: hun
3 @Participants: CH1 Target_Child, FW1 Ki Investigator
4 @ID: hun|HUKILC|CH1|4;6.|female|ak||Target_Child||
5 @ID: hun|HUKILC|FW1||female|||Investigator||
6 @Date: 03-FEB-2012
7 @Media: hf_5, audio
8 @Transcriber: Juli
9 @Transcription: anonymized
10 @Situation: II./XII. kerületi óvodában készült a felvétel, a
11 szobába beszűrődik háttérzaj.
12 *FW1: na .
13 %arg:
14 *FW1: akkor mondd meg nekem légyszíves , hogy hívnak .
15 *CH1: Horváth Márta Ilona .
16 *FW1: mhm .
17 *FW1: és hány és éves vagy Márta ?
18 *CH1: négy és fél .
19 *FW1: mhm .
20 *FW1: és mi a beceneved ?
21 *CH1: nem tudom .
22 %arg: $SUBJ:Exper=human+nom&lit|verb
23 *FW1: nem tudod ?
24 *FW1: mit szokott anya mondani , meg a barátaid ?
25 *FW1: Márta ?
26 *FW1: vagy azt , hogy Márti ?
27 *FW1: mit ?
28 *CH1: litkán [: ritkán ] [*] azt szokták mondani köszönésnek a barátaim
29 [: barátaim] [*] , hogy szia Márta .
30 *CH1: szia mm@fp .
18oct18[C|CHAT] * 17
$SUBJ $COMPL
$OBJ
$IO
$OBL
```

CLAN

CLAN

- CHAT mode:
 - Tierekkel segíti az átírást
 - CHAT szintaxist ellenőrzi a depfile.cut file-ban megadott szabályok alapján
- Sonic mode: Segít szinkronizálni az átírást és hangot/videót.
- Coder mode:
 - Kódolást egyszerűsíti. Külön file-ban elmentett, hierarchikusan rendezett kódrendszer hív be.

Elemző programok

- Cmd/ Ctrl d a CLAN-ból vagy a neten is használható
- Working directory: az adatok mappája
- Lib: a CLAN .cut file-ok
- Progs: a különböző elemzőprogramok (freq, kwal, maxwd, mlu)

név - segítség

név +t*BESZÉLŐ +t%altier +s"string" fileok*.cha

Egyéb: LuCiD toolkit

- CHILDES Browser: Célrányos keresés korpuszméret és kor szerint
- CHILDES Generator: megbecsüli egy kifejezést követő szavak valószínűségét a korpusz alapján
- Distributional Word Classification: klaszterekbe rendez szavakat a disztribúciós tulajdonságaik szerint (milyen szavakkal fordulnak elő)