Carnegie Mellon University

TalkBankDB: A Comprehensive Data Analysis Interface to TalkBank

John Kowalski, Brian MacWhinney

About TalkBank

TalkBank, a **CLARIN B Centre**, is the host for a collection of multilingual multimodal corpora designed to foster fundamental research in the study of human communication.

It contains tens of thousands of audio and video recordings across many languages linked to richly annotated transcriptions, all in the CHAT transcription format.

The origins of TalkBank trace back to 1984 with the creation of the CLAN (Child Language Anlaysis) tools and the associated CHAT transcription format.

The corpus began with annotated media of child language acquisition (CHILDES database) and has expanded to include fourteen annotated media language databases including:

- SLABank for studying second-language acquisition.
- CABank for conversational data.
- ClassBank for study of language in the classroom.
- SamtaleBank for the study of Danish conversations.
- Along with a series of clinical databanks for aphasia, stuttering and other disorders.

The size and scope of TalkBank continues to expand. As of this writing, TalkBank includes over 5TB of richly annotated media.

TalkBank Usage

	CHILDES	AphasiaBank	PhonBank	FluencyBank	HomeBank	TalkBank
Age (years)	30	10	7	1	2	14
Words (millions)	59	1.8	0.8	0.5	audio	47
Linked Media (TB)	2.8	0.4	0.7	0.3	3.5	1.1
Languages	41	6	18	4	2	22
Publications	7000+	256	480	5	7	320
Users	2950	390	182	50	18	930
Web hits (millions)	5.0	0.5	0.1	0.1	0.4	1.7

TalkBankDB Goals

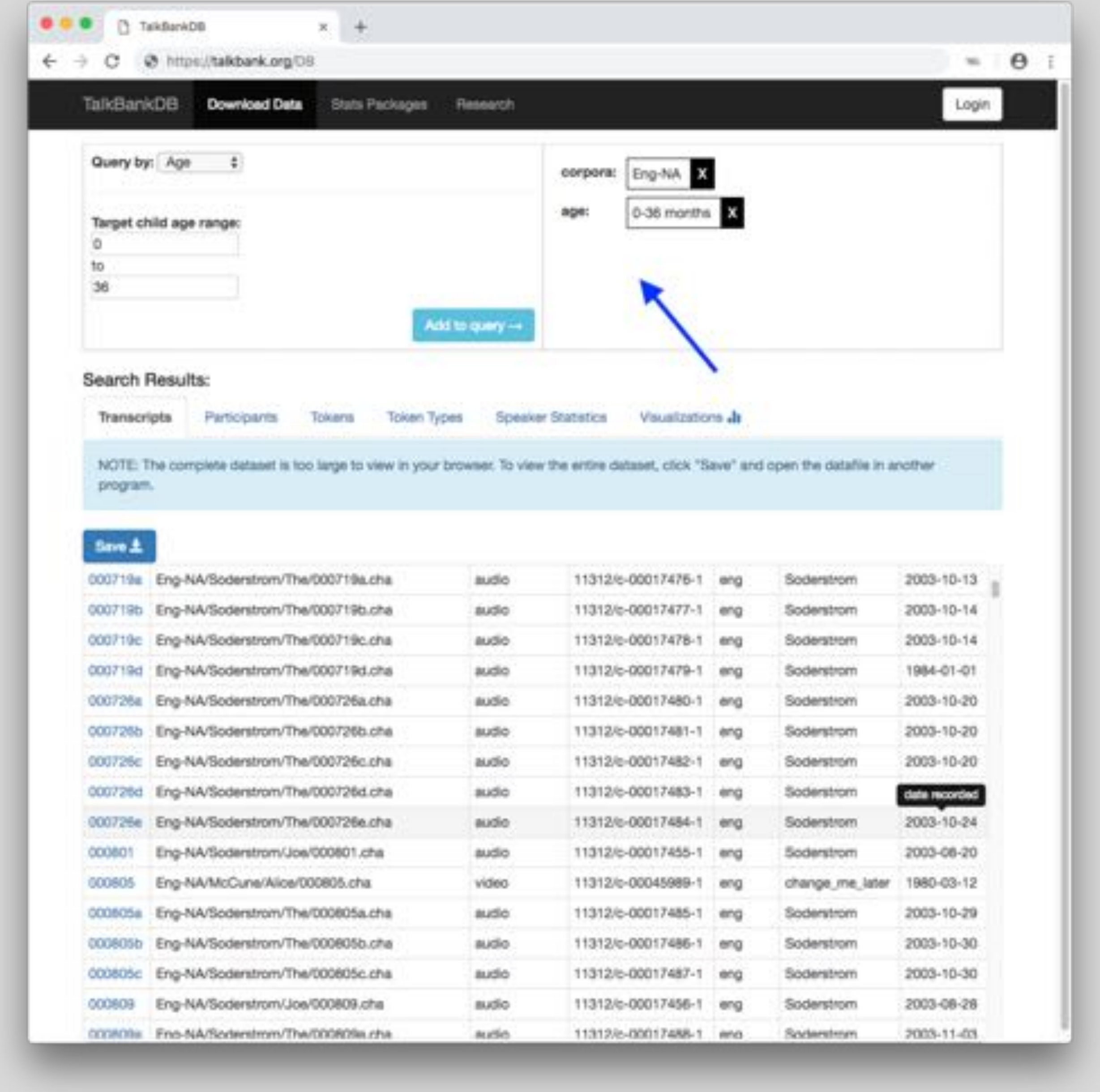
The purpose of TalkBankDB is to provide an intuitive on-line interface for researchers to explore TalkBank's media and transcripts, specify data to be extracted, and pass these data on to statistical programs for further analysis. We hope to:

Increase accessibility:
Previously, browsing TalkBank required knowing the name of a corpus or area of research, finding its location within the talkbank.org domain (ex: fluency.talkbank.org), then browsing/downloading the media and annotations and installing the CLAN tools. To make these resources more accessible, TalkBankDB provides a single online interface to query across all of TalkBank to find relevant corpora and links to media and transcripts. Users can visually explore data directly in the browser, and if desired, download it for further analysis in a statistical software package.

Grow community and foster collaboration:
With the entirety and richness of TalkBank freely accessible from a simple web interface, resources that were previously known only by advanced users will be open to a broader community. Features such as word usage, utterance length, measures of language acquisition speed and ability by demographics can easily be selected, output, plotted, and analyzed through the web interface. By also providing a GitHub account link for users to upload scripts and analyses, the TalkBankDB site provides a single point where users can explore, share their research, and see what others are doing in the TalkBank community.

Example TalkBankDB WorkFlow

Build Query to Database



Query by: Age

Target child age range:
0
to
36

corpora: Eng-NA X

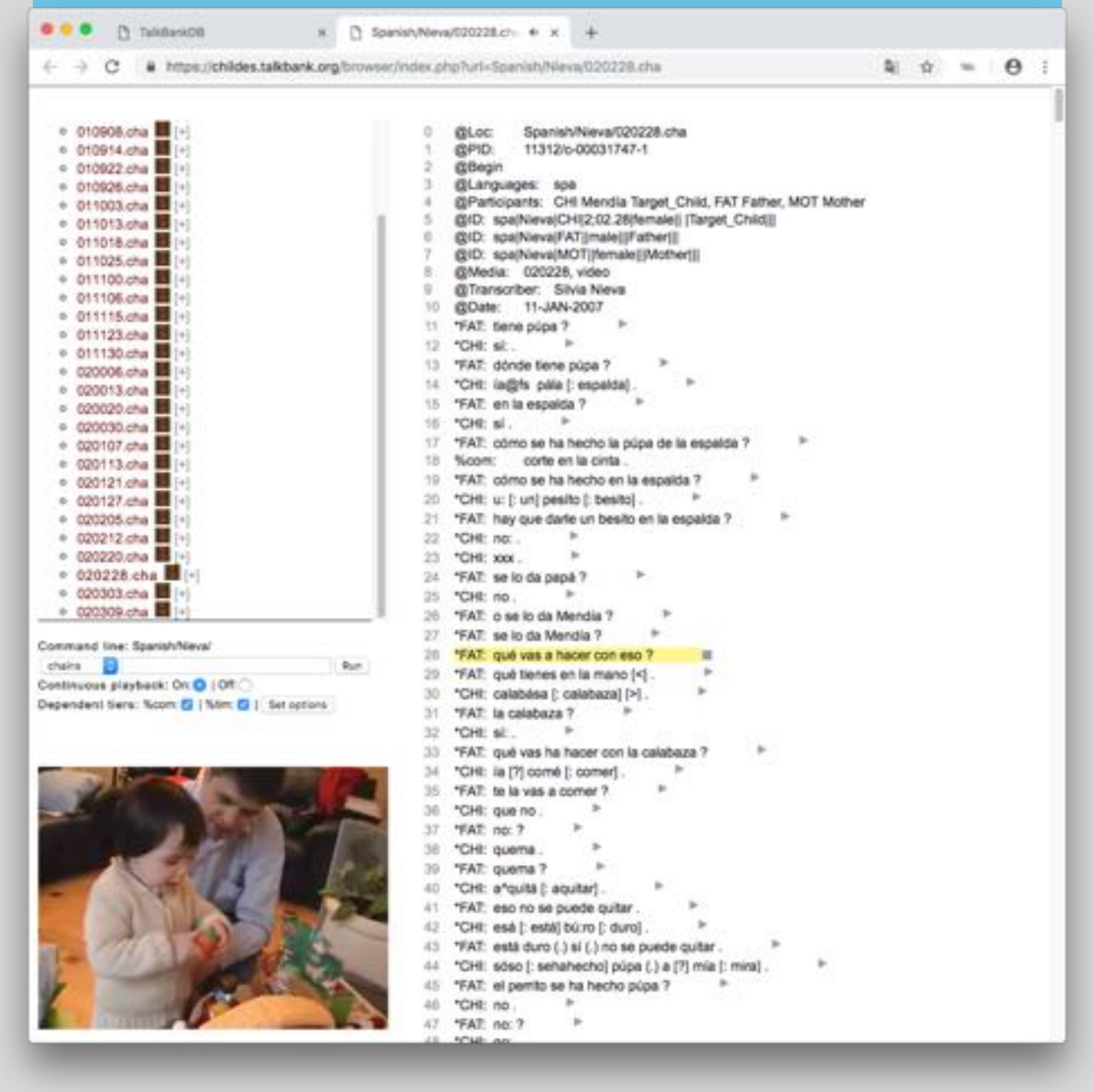
age: 0-36 months X

Search Results:

NOTE: The complete dataset is too large to view in your browser. To view the entire dataset, click "Save" and open the datafile in another program.

000719a	Eng-NA/Soderstrom/The/000719a.cha	audio	11312/c-00017476-1	eng	Soderstrom	2003-10-13
000719b	Eng-NA/Soderstrom/The/000719b.cha	audio	11312/c-00017477-1	eng	Soderstrom	2003-10-14
000719c	Eng-NA/Soderstrom/The/000719c.cha	audio	11312/c-00017478-1	eng	Soderstrom	2003-10-14
000719d	Eng-NA/Soderstrom/The/000719d.cha	audio	11312/c-00017479-1	eng	Soderstrom	1984-01-01
000726a	Eng-NA/Soderstrom/The/000726a.cha	audio	11312/c-00017480-1	eng	Soderstrom	2003-10-20
000726b	Eng-NA/Soderstrom/The/000726b.cha	audio	11312/c-00017481-1	eng	Soderstrom	2003-10-20
000726c	Eng-NA/Soderstrom/The/000726c.cha	audio	11312/c-00017482-1	eng	Soderstrom	2003-10-20
000726d	Eng-NA/Soderstrom/The/000726d.cha	audio	11312/c-00017483-1	eng	Soderstrom	date recorded
000726e	Eng-NA/Soderstrom/The/000726e.cha	audio	11312/c-00017484-1	eng	Soderstrom	2003-10-24
000801	Eng-NA/Soderstrom/Joe/000801.cha	audio	11312/c-00017455-1	eng	Soderstrom	2003-08-20
000805	Eng-NA/McCune/Alice/000805.cha	video	11312/c-00045989-1	eng	change_me_later	1980-03-12
000805a	Eng-NA/Soderstrom/The/000805a.cha	audio	11312/c-00017485-1	eng	Soderstrom	2003-10-29
000805b	Eng-NA/Soderstrom/The/000805b.cha	audio	11312/c-00017486-1	eng	Soderstrom	2003-10-30
000805c	Eng-NA/Soderstrom/The/000805c.cha	audio	11312/c-00017487-1	eng	Soderstrom	2003-10-30
000809	Eng-NA/Soderstrom/Joe/000809.cha	audio	11312/c-00017456-1	eng	Soderstrom	2003-08-28
000809a	Eng-NA/Soderstrom/The/000809a.cha	audio	11312/c-00017488-1	eng	Soderstrom	2003-11-03

Click Link in Results to Play Media and View Transcripts

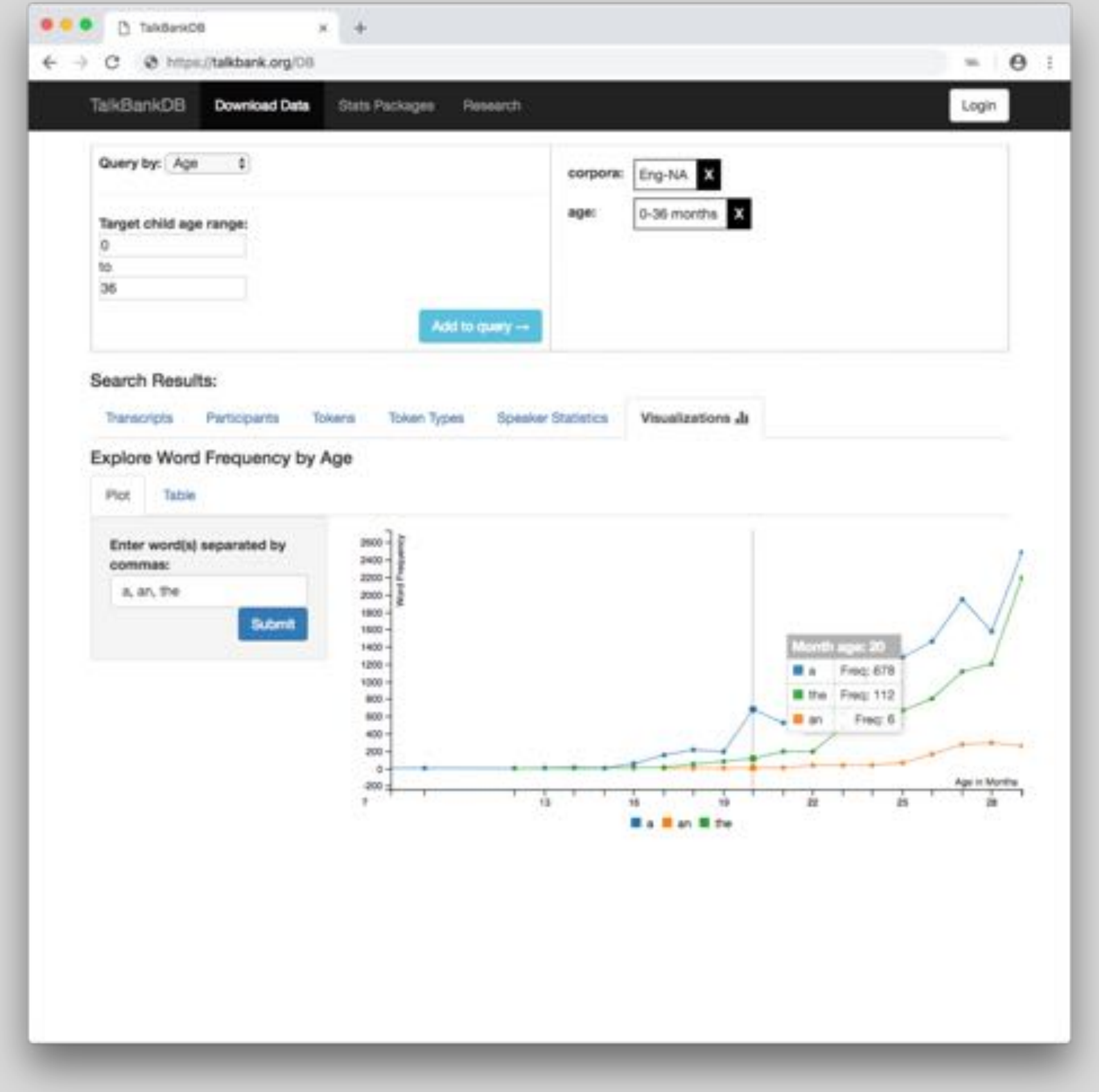

TalkBankDB
Spanish/Nieva/020228.cha
https://childes.talkbank.org/browser/index.php?url=Spanish/Nieva/020228.cha
Run
Set options

010908.cha [+]
010914.cha [+]
010922.cha [+]
010926.cha [+]
011003.cha [+]
011013.cha [+]
011018.cha [+]
011025.cha [+]
011100.cha [+]
011106.cha [+]
011115.cha [+]
011123.cha [+]
011130.cha [+]
020006.cha [+]
020013.cha [+]
020020.cha [+]
020030.cha [+]
020107.cha [+]
020113.cha [+]
020121.cha [+]
020127.cha [+]
020205.cha [+]
020212.cha [+]
020220.cha [+]
020228.cha [+]
020303.cha [+]
020309.cha [+]

Command line: Spanish/Nieva/
chains
Continuous playback: On ● | Off ○
Dependent tiers: %com: ☑ | %tim: ☑ |

```
@Loc:      Spanish/Nieva/020228.cha
@PID:      11312/c-00031747-1
@Begin
@Languages:   spa
@Participants:   CHI Mendia Target_Child, FAT Father, MOT Mother
@ID:  spa|Nieva|CHI|2;02.28|female|||Target_Child|||
@ID:  spa|Nieva|FAT||male|||Father|||
@ID:  spa|Nieva|MOT||female|||Mother|||
@Media:   020228, video
@Transcriber:   Silvia Nieva
@Date:    11-JAN-2007
*FAT:  tiene púpa ?
*CHI:  sí: .
*FAT:  dónde tiene púpa ?
*CHI:  ia@fs  pàla [: espalda] .
*FAT:  en la espalda ?
*CHI:  sí .
*FAT:  cómo se ha hecho la púpa de la espalda ?
%com:     corte en la cinta .
*FAT:  cómo se ha hecho en la espalda ?
*CHI:  u: [: un] pesito [: besito] .
*FAT:  hay que darle un besito en la espalda ?
*CHI:  no: .
*CHI:  xxx .
*FAT:  se lo da papá ?
*CHI:  no .
*FAT:  o se lo da Mendia ?
*FAT:  se lo da Mendia ?
*FAT:  qué vas a hacer con eso ?
*FAT:  qué tienes en la mano [<] .
*CHI:  calabésa [: calabaza] [>] .
*FAT:  la calabaza ?
*CHI:  sí: .
*FAT:  qué vas ha hacer con la calabaza ?
*CHI:  ia [?] comé [: comer] .
*FAT:  te la vas a comer ?
*CHI:  que no .
*FAT:  no: ?
*CHI:  quema .
*FAT:  quema ?
*CHI:  a*quitá [: aquitar] .
*FAT:  eso no se puede quitar .
*CHI:  esá [: está] bú:ro [: duro] .
*FAT:  está duro (.) sí (.) no se puede quitar .
*CHI:  sóso [: senahecho] púpa (.) a [?] mía [: mira] .
*FAT:  el perrito se ha hecho púpa ?
*CHI:  no .
*FAT:  no: ?
```

Explore Data with Visualizations



Query by: Age

Target child age range:
0
to
36

corpora: Eng-NA X

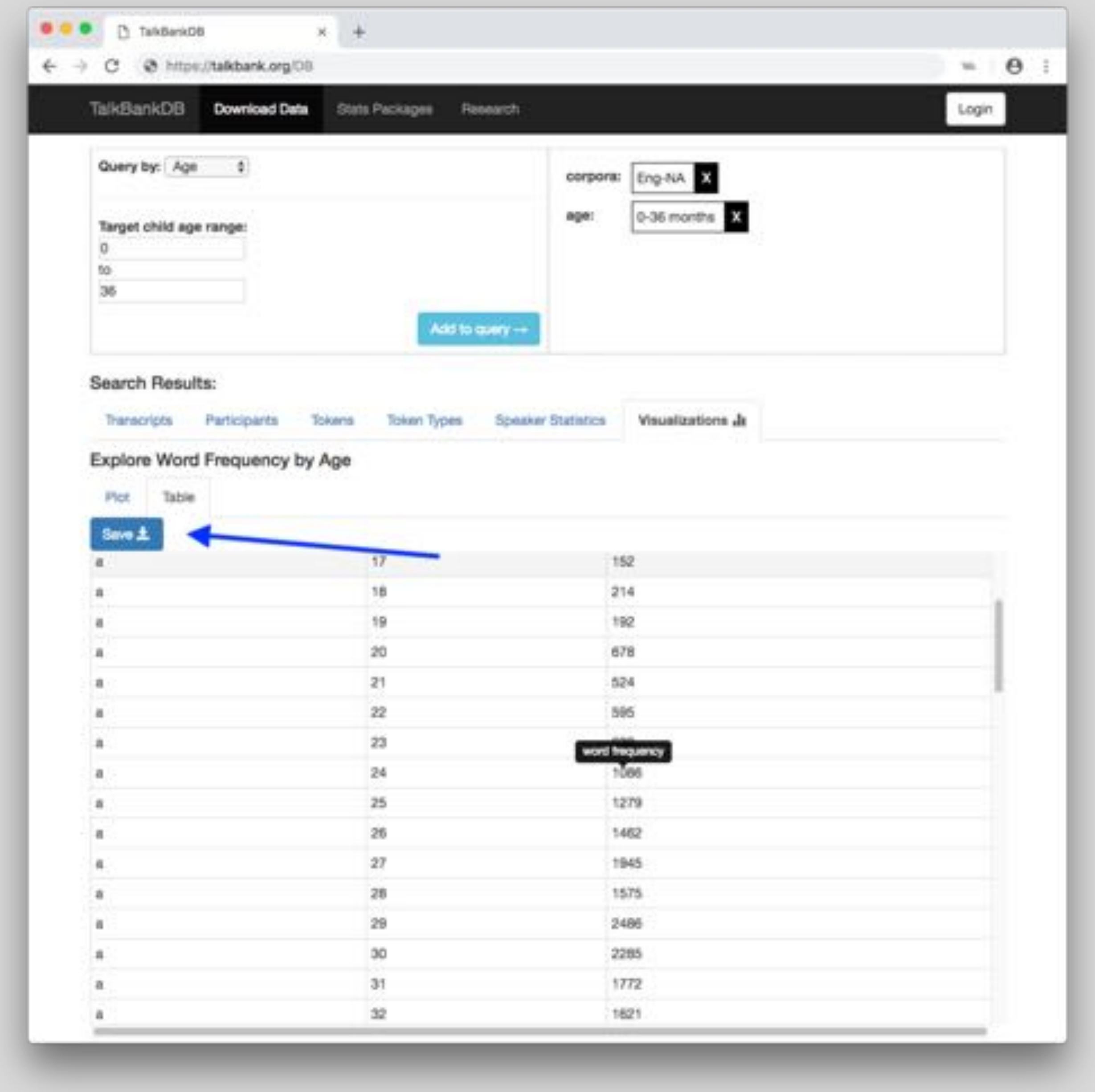
age: 0-36 months X

Search Results:

Explore Word Frequency by Age

Enter word(s) separated by commas:
a, an, the

Save Data for Import to Statistical Software (R/NumPy/…)



Query by: Age

Target child age range:
0
to
36

corpora: Eng-NA X

age: 0-36 months X

Search Results:

Explore Word Frequency by Age

a	17	152
a	18	214
a	19	192
a	20	678
a	21	524
a	22	595
a	23	[illegible]
a	24	1086
a	25	1279
a	26	1462
a	27	1945
a	28	1575
a	29	2486
a	30	2285
a	31	1772
a	32	1621

Technical Details

Creation of the TalkBankDB database relies on the fact that all TalkBank transcripts are pure UTF-8 text files that explicitly implement the CHAT annotation format. These files are then processed by the CHATTER Java program. CHATTER can convert a CHAT file to XML that can be round-tripped back to the file's original CHAT format. The XML format and the associated schema facilitates use of TalkBank corpora by third party programs and systems, eliminating the need to parse complex raw strings.

Since JSON can be used directly by front-end web apps, we eliminate the need of the app to constantly convert XML to JSON and back again by first converting the XML transcripts outputted by CHATTER to JSON using xml-js (Nashwaan, 2018). This tool supports bidirectional XML/JSON conversion. So, combined with CHATTER, we can round-trip from JSON to the original CHAT formatted transcript.

To store our collection of JSON documents, we use MongoDB, a widely-used free and open-source document database. An added benefit of this document database is it makes scaling to increasing data demands easy by allowing the database to scale out across multiple inexpensive machines through "sharding" of the database. This can be difficult to do with relational databases, where often the only option is to "scale up" by purchasing increasingly powerful machines. The scaling-up strategy is not always possible, and can one day be unable to meet the growing size and computational demands of the database.

The front end web interface is written in standard HTML, CSS, and JavaScript to ensure cross-browser support. Visualizations are through the C3.js library, although many others can be used. Care is taken so the JavaScript code is clearly commented and maintainable, following the popular "web component" design pattern common in many large-scale web apps.

Future Work

Started only this year, TalkBankDB is still in its early stages. Development is very active, so there are many new features coming soon, including:

Query builder:

- Currently, the query builder ANDs between query categories (corpus, language, media type, gender, age) and ORs within categories. We are working on a UI allowing users to build any logical combination of features (AND, OR, NOT) with expressivity similar to the Corpus Query Language (CQL).
- Additionally, we will expand beyond the current five query categories (ex: morphological properties).

UI to Add/Edit documents with DB versioning:

- This will necessitate authenticating users to authorize DB modifications.
  - Additionally, each download will include a version number for reproducibility of analyses.

Additional Visualizations and Analyses:

- The code for each visualization is encapsulated in its own JavaScript module (file).
- Many other visualizations can be made following this module's pattern.
  - Users can contribute interesting visualizations.

Direct access to TalkBankDB from statistical software:

- As a convenience, we plan to make available R and Python libraries for querying and downloading data from TalkBankDB directly, without having to go through the TalkBankDB web interface, download, then include.

References

[Chang 2017] Chang, F. (2017) The LuCiD language researcher's toolkit [Computer software]. Retrieved from http://www.lucid.ac.uk/resources/for-researchers/toolkit/

[MacWhinney 2000] MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates

[MacWhinney 2014] MacWhinney, B. (2014). The childes project: Tools for analyzing talk, volume ii: The database. Psychology Press.

[Nashwaan 2018] Nashwaan, Yousuf, xml-js, (2018) GitHub repository, https://github.com/nashwaan/xml-js

[Sanchez] Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. C. (in prep). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System (CHILDES). Manuscript in preparation.

Acknowledgements

We would like to acknowledge support from these grants: NIDCD AphasiaBank DC008524, NICHD PhonBank HD051698, NICHD CHILDES HD082736, NSF HomeBank SMA-1539010, NIDCD FluencyBank DC015494, and NSF FluencyBank BCS-1626294.