TalkBankDB: A Comprehensive Data Analysis Interface to TalkBank

John Kowalski
Carnegie Mellon University, USA
jkau@andrew.cmu.edu

Brian MacWhinney
Carnegie Mellon University, USA
macw@andrew.cmu.edu

Abstract

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 purpose of the TalkBankDB project 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.

1 Introduction

The origins of TalkBank trace back to 1984 with the creation of the CLAN (Child Language Analysis) tools and the associated CHAT transcription format (MacWhinney, 2000). 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, and 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 8TB of annotated media.

	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

Table 1: TalkBank Usage

Currently, most users interact with TalkBank data by using the CLAN program. CLAN consists of a set of tools for transcribing media in CHAT format, playing back the media with time-stamped annotations, and extracting statistics and metadata from a set of transcripts. CLAN has been refined for decades and is highly capable. However, using it requires a significant effort from the researcher to study the CLAN manual and learn CHAT annotations. Moreover, CLAN is mostly tuned for creating new transcripts and for working with single corpora. It is not designed for systematic queries of the entire TalkBank database to extract general patterns and statistics. Because of these limitations, CLAN is not an ideal tool for researchers who want to conduct wider corpus analyses on the existing database. Here we report on a new system, called TalkBankDB, designed to provide this additional functionality.

2 Increasing the Accessibility of TalkBank Corpora

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), and then browsing/downloading the media and annotations.

Without prior knowledge of how TalkBank is structured and what corpora exist within each TalkBank collection, users may be unaware that particular resources exist. TalkBankDB provides a single online interface to query across all of TalkBank to find the names and categories of relevant corpora and links to media and transcripts. For example, a query for the Spanish language yields a list of transcripts within TalkBank spanning many separate corpora. Further queries can limit by date of recording, native language of speakers, age of participants, media type (audio/video/none), and others. The user will then have a list of all media and descriptive metadata matching their query, with links to each directly playable from the browser. After a query is submitted, clickable tabs appear to show descriptive lists of participants in matched transcripts, word tokens spoken, tokens grouped by type, and statistics for each speaker (number of words spoken, mean utterance length, and others.) TalkBankDB allows users to construct new combinations of corpora or subparts of corpora based on features they define (Figure 1 and Figure 2).

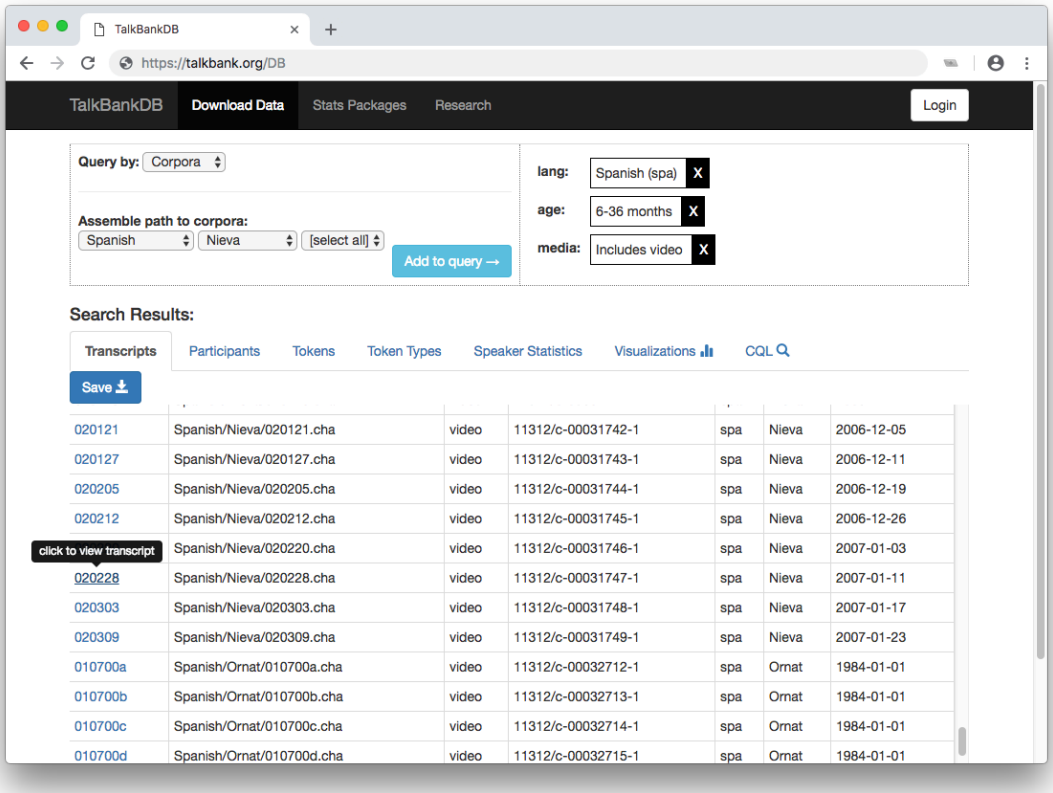

Figure 1. A query yields a table of all matching documents with metadata for each, allowing the user to further refine the query.

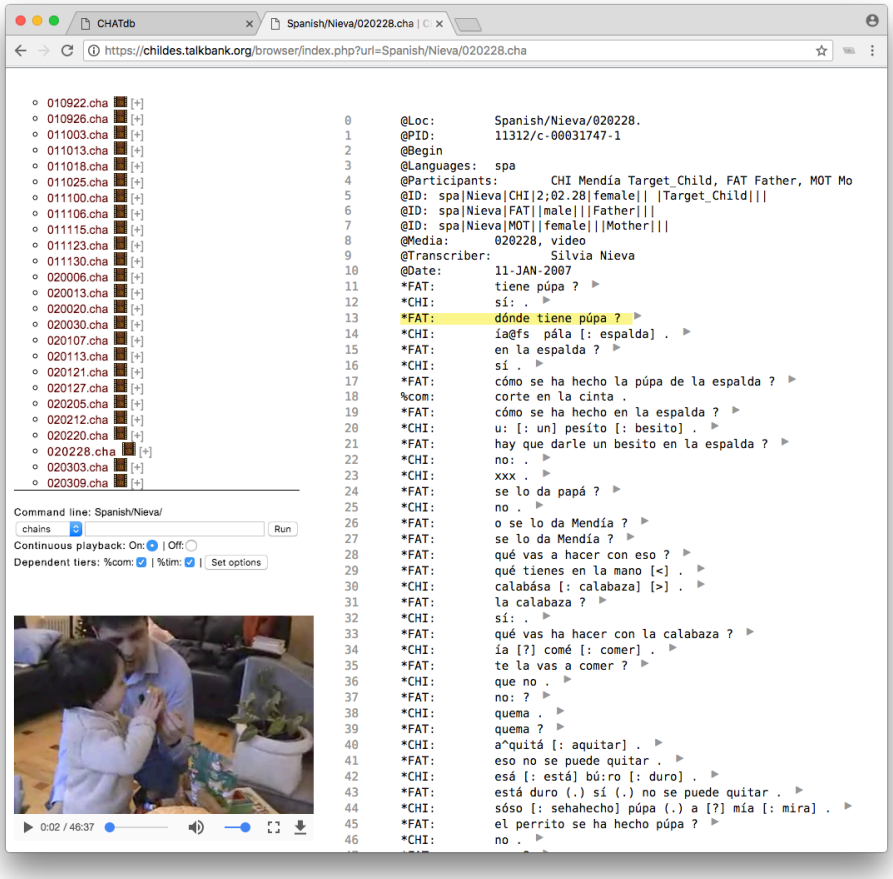

Figure 2. Clicking on the name of the transcript loads the corresponding media and annotations in the browser and allows for direct playback of the media.

In addition to using TalkBankDB to locate transcripts and media with specific features across TalkBank, researchers can derive statistical summaries of the annotations in the transcripts. A pulldown of variables to be extracted includes the age range of participants, the roles of speakers (mother, father, child, teacher, etc), the number of words spoken, mean utterance length, specific words used, and others. For instance, one can make a plot of frequencies of English article usage (a/an, the) by mothers speaking to their children in relation to their child's age. The exploration space enabled by this simple interface is huge.

Child language researchers had already built two systems designed to achieve this type of functionality. These are the childes-db project (Sanchez et al., 2018) and the LuCiD Toolkit (Chang, 2017). Both of these projects were created to analyse only the portion of TalkBank dealing with child language acquisition (CHILDES corpus), whereas TalkBankDB encompasses the whole of TalkBank. The principle goal of these systems is to output spreadsheets which can then be passed on to statistical analysis by systems such as R, NumPy, or Excel. TalkBankDB also provides this functionality.

The childes-db project (childes-db.stanford.edu) offers both a web interface and R package to analyse CHILDES. Downloaded CHILDES data are stored in a MySQL database. There are six main functions in the childes-db R library: get_transcripts(), get_participants(), get_tokens(), get_types(), get_utterances(), and get_speaker_statistics(). For the web interface, childes-db employs R Studio's Shiny Server enabling the plotting of variables also accessible from the aforementioned R library functions. The LuCiD Toolkit offers similar facilities to childes-db for exploring the CHILDES corpus. It also employs a Shiny server (gandalf.talkbank.org:8080) to offer a web interface to extract and analyse variables from the transcripts. However, this facility is based on a 140GB spreadsheet set that contains data and precomputed statistics from CHILDES. TalkBankDB differs from these facilities by creating

an editable document database from which statistics are computed dynamically, creating a more scalable and flexible system.

3 Database Architecture and Implementation 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, available from https://talkbank.org/software/chatter.html. 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, TalkBankDB eliminates the need for 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, the system can provide a verifiable round-trip from JSON to the original CHAT formatted transcript.

Since much of the data and metadata contained within the TalkBank CHAT transcripts are subject to change and amplification, using a relational database like mySQL with a strict tabular schema is not as practical as a more flexible document database. The effort to pre-set a fixed schema with a normalized relational database can cause problems when the schema needs to be modified and extended with new phonology, sequence numbers for tiers, adding TEI annotations, etc.

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 that it makes it easy to scale up to increasing data demands by allowing the database to be encoded across multiple inexpensive machines through "sharding". This can be very difficult to do with relational databases, where often the only option is to "scale up" by purchasing increasingly powerful machines. The strategy of scaling up through use of a single larger machine is not always possible, and it may eventually 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. 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.

Initially, TalkBankDB will include only publicly accessible data. Access will be controlled by the CLARIN single sign-on authentication system. Access to private clinical data will require a second-level of authentication.

4 Database JSON Format

The JSON format derived from a CHAT file (via CHATTER XML) is very simple and extensible. An example JSON representation of the utterance "talking to the tape recorder" is below:

```
{
  "who":"FAT",

  "uID":"u8",

  "words":[
    {
       "w":"talking",
       "mor":{
          "stem": "talk",
          "pos": "part"
       }
    },
    {
       "w":"to",
       "mor":{
```

```
            "stem": "to",
            "pos": "prep"
         }
      },
      {
         "w":"the",
         "mor":{
            "stem":"the",
            "pos": "det"
         }
      },
      {
         "w":"tape",
         "mor":{
            "stem":"tape",
            "pos": "n"
         }
      },
      {
         "w":"recorder",
         "mor":{
            "stem":"record",
            "pos": "n"
         }
      }
   ],
   "media":{
      "start": 20.062,
      "end": 20.805"
   }
}
```

Here we see that for each utterance:

- A speaker is defined, here as the father (`who: "FAT"`).
- The utterance's sequence number within the transcript (`uID: "u8"`).
- An array of words (`words: []`).
- A start/end time of the recording (`media: {start: 20.062, end: 20.805}`).

Each word object consists of:

- A word as it appears in the recording (`w: "talking"`).
- A morphology object consisting of an extensible number of properties. (`mor: {}`).
    - Here we define:
        - The stem or lemma of the word (`stem: "talk"`).
        - Part of speech of the word, here participle (`"pos: "part"`).
        - Other key/value pairs.

This simple utterance object is the fundamental building block of the JSON representation of a CHAT transcript and thus of TalkBankDB. All tab-delimited data downloads, statistics, visualizations, and token/grammatical pattern searches mentioned in this manuscript are derived from this repeating structure.

As of this writing, TalkBankDB is still less than a year old and we are in the process of adding additional information regarding specific coding features in CHAT transcripts for the %pho phonology line, the %mor morphology line, and the %gra grammatical relations line . However, the fundamental JSON-based structure will stay the same. Corpora outside TalkBank that have words tagged with stem and part of speech could also be converted to this simple format. They would then immediately get the benefits of the analysis, visualization, and search tools of TalkBankDB.

5 Searching for Token and Grammatical Patterns

TalkBankDB provides a toolkit to search for token and grammatical structures across TalkBank corpora. The toolkit interface is a language composed of a combination of regular expressions and a syntax to specify patterns of grammatical and other semantic tags. This query language is based on the popular Corpus Query Language (CQL) used by SketchEngine software (www.sketchengine.eu) (Kilgarriff, 2004), which in turn is based on the query language of the Corpus Work Bench (CWB) project developed at the University of Stuttgart (Christ, 1994). Since CQL syntax is familiar to many researchers, TalkBankDB implements a large subset of it.

In CQL, searching for a token takes the form of:
`[attribute="value"]`

Where `attribute` is often one of:

- "`word`" to match a particular word or set of words defined with a regular expression.
- "`lemma`" to match all forms of a lemma. Ex: jump → jump, jumps, jumping, jumped.
- "`pos`" to match a word based on its part of speech tag.

Any attribute/value defined on a word in the database can be searched this way. The full search language grammar along with all attribute/value pairs are defined on the TalkBankDB website. Listed below are examples of some common CQL search patterns.

To find all instances of...

- The word "bring":
  `[word="bring"]`

- Verb forms of "bring" (bring, brings, bringing, brought):
  `[lemma="bring"]`

- The sequence "bring your hat":
  `[word="bring"] [word="your"] [word="hat"]`

- The sequence "bring your" followed by a noun:
  `[word="bring"] [word="your"] [pos="N"]`

- The sequence "bring your" followed by an adjective and noun:
  `[word="bring"] [word="your"] [pos="ADJ"] [pos="N"]`

- The sequence "bring your" followed by one or more adjectives and a noun:
  `[word="bring"] [word="your"] [pos="ADJ"]+ [pos="N"]`

- The sequence "bring a/an/the" followed by one or more adjectives and a noun:
  `[word="bring"] [word="an?|the"] [pos="ADJ"]+ [pos="N"]`

To search for keyword/grammatical patterns in TalkBankDB, first define the corpus to search (by name, language, media present, etc.), then click on the "CQL" tab, enter a CQL query in the text field, and click "Submit". This generates a list of "concordances", matched keywords along with the words surrounding each. The concordances listed are in the format of a "Key Word in Context" (KWIC), where the matched keywords are highlighted in blue surrounded by the text that contains it in the transcript. In addition, the transcript path, speaker, and utterance ID within the transcript are also listed to provide additional context (Figure 3).

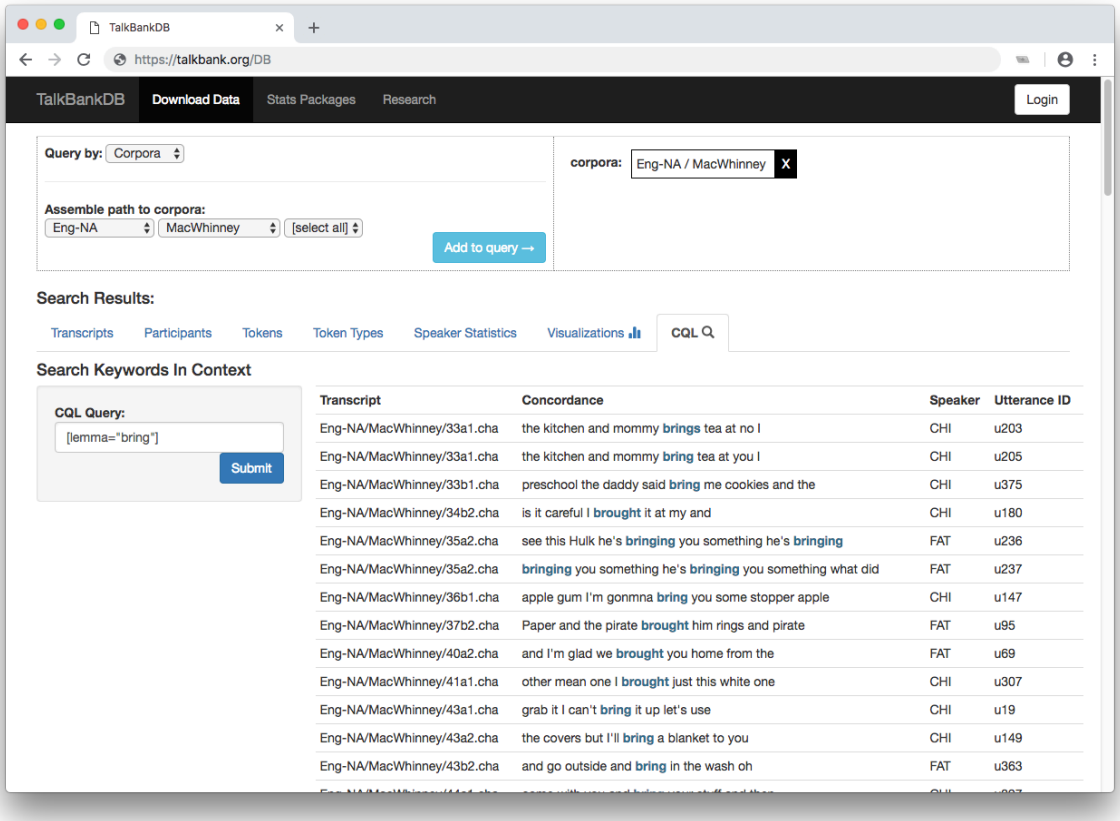

Figure 3. Selecting the MacWhinney corpus, then choosing "CQL" tab to generate KWIC concordances for the lemma of verb "bring".

6 Visualizations

TalkBankDB also provides methods for client-side data visualization. After selecting a corpus, clicking on the "Visualizations" tab opens a collection of options for generating plots and for downloading the data used to generate these plots. The layout for each visualization UI is similar. On the left is a UI for specifying the options and information needed to generate a plot; on the right is the interactive plot generated from this information. At the top is the default "Plot" tab just described and a "Table" tab that displays the downloadable data for the current plot. Figure 4 shows the "Word Frequency by Age" visualization for the MacWhinney corpus, choosing the English articles (a, an, the).

In contrast to servers like the RStudio Shiny Server (shiny.rstudio.com) that process data for creation of a static image to be displayed in the browser, TalkBankDB's server simply sends back data, thereby giving the client freedom to plot with any visualization library. Many visualization libraries are freely available today, such as Chartist, Highcharts, Google Charts, Chart.js, and others. New libraries can be swapped in and out of TalkBankDB as required. We chose the open-source C3.js library, because it is easy to use and covers all the plots we currently need. Moreover, the plots it generates can be explored by zooming in/out and dragging along values of the x-axis. Another benefit of C3 is that it is based on D3, a very powerful graphical toolkit for browsers. If a visualization is desired but not covered by C3, one can extend the code using the powerful toolkit that D3 provides.

TalkBankDB currently provides visualizations for exploring word frequency by age, number of utterances/words by age, mean word length of utterances (MLUw), and type-token ratio (TTR). Adding more visualizations to TalkBankDB is straightforward. Each client-side visualization is processed by its own JS module that follows a 3-step pattern of:

1. Make API request to server to retrieve data on selected corpus.
2. Process or compute stats on data.
3. Call a visualization function to plot processed data.

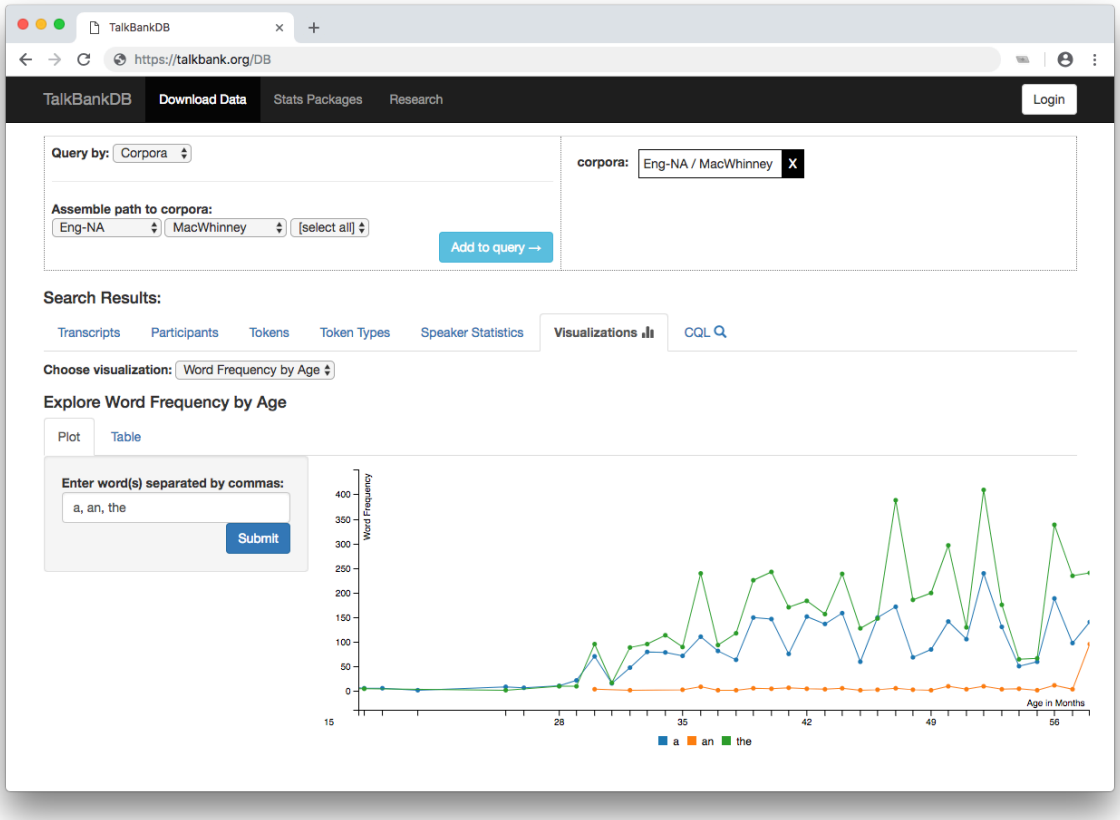

Figure 4. Plotting word frequency by age of English articles (a, an, the) by age within the MacWhinney corpus.

7 Other Features

A beta version of TalkBankDB is currently at https://talkbank.org/DB. The features offered will be refined and expanded on the basis of input from the extended CLARIN community. Below we list some features in the current beta specification:

- Button to download local copies of tab-delimited tables generated by TalkBankDB queries for use in further statistical analysis.
- Include links in tables returned by queries to open and play audio/video transcripts in browser.
- Option to upload new files, define new TalkBank corpora branches.
- Option to view/edit transcripts.
- Maintain state in state of user's queries and analyses in URL so that analyses can be shared with others by sending a unique URL.

8 Related Work

The design and scope of TalkBankDB has been influenced by our work with several related projects, including SketchEngine (sketchengine.eu), Corpus Workbench (cwb.sourceforge.net), EXMARaLDA (exmaralda.org), MTAS (meertensinstituut.github.io/mtas), ANNIS (corpus-tools.org/annis), and Alpheios (alpheios.net), as well as the childes-db and LuCiD Toolkit projects mentioned earlier. These systems have features that continue to influence the development of TalkBankDB, especially for facilitating the creation and interpretation multi-tier annotations of multimedia corpora.

9 Additional Applications and Expansions

Although TalkBankDB is designed around the CHAT format, it can be applied to other formats and projects in the CLARIN ecosystem. Since the format stored in TalkBankDB is not CHAT, but a simplified JSON representation, including documents in TalkBankDB only requires a script to convert from another (non-CHAT) CLARIN format to this JSON format. The JSON schema currently includes entries for metadata such as document name, version number, corpus name, and media type. In addition, it has a list of participants, and an "utterances" array with an entry for each word, with each word supplemented with metadata including speaker ID, token morphology, and utterance number. Any format encoding transcripts of spoken text with morphological tagging can easily be adapted for inclusion in TalkBankDB.

10 Conclusion

A main goal of TalkBankDB is to provide the CLARIN/TalkBank community with easier access to TalkBank data and analysis. 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. The TalkBankDB interface can also be used in classroom demonstrations and project assignments for humanities or data analysis students, increasing awareness of the CLARIN community and inspiring future members.

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/

[Christ 1994] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *arXiv preprint cmp-lg/9408005*.

[Kilgarriff 2004] Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell. Itri-04-08 the sketch engine. Information Technology, 2004.

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

MongoDB [Computer software]. (2018). Retrieved from https://www.mongodb.com.

Node.js [Computer software]. (2018). Retrieved from https://nodejs.org/en.

[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. (2019). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System (CHILDES). Behavior Research Methods. 1-14.