LUCS Minor 16, 2010. ISSN 1104-1609.

ChildFreq: An Online Tool to Explore Word Frequencies in Child Language

Rasmus Bååth
Lund University Cognitive Science
Kungshuset, Lundagård, 222 22 Lund
rasmus.baath@lucs.lu.se

Abstract

This technical report describes the implementation and use of ChildFreq, a tool for assessing lexical norms of children from one to seven years old. As the name implies, ChildFreq works by extracting word frequencies from a large corpus of child language. These can then be ordered by age or mean length of utterance, and it is also possible to split the data by the children's gender. A query of words to count the frequency of produces both a line chart and a table with more detailed information. The child language data is taken from the English part of the CHILDES database[1] and comprises more than 5,000 transcriptions,a total of $\approx 3,500,000$ word tokens. The children's ages range from six months to seven years, with most children being three years old. ChildFreq is freely available online at `http://childfreq.sumsar.net` .

Keywords: language development, word frequency, CHILDES , lexical acquisition, lexical norms.

1 Introduction

It is truly an amazing feat for a child to acquire a language. From the age of one to the age of seven, a child goes from knowing just a few words to having a vocabulary of roughly 10,000 words. Acquiring a vocabulary does not happen at an uniform rate, either in respect to the distribution of when words are learned, or in what order they are learned. Starting from around the age of one, an average child learns fifty to a hundred words within six months. Word learning accelerates after that so that by the age of two-and-a-half, a child uses slightly more than 500 words (Goodman et al., 2008). Of these, most will be concrete nouns. It is known that concrete nouns are usually learned before more abstract ones and that nouns in general are learned earlier than verbs (Golinkoff and Hirsh-Pasek, 2008). How language is acquired not only is of relevance to understanding child development but also touches on issues of language usage more generally, as well as concept formation.

When studying language in children, it is useful to have information regarding children's *lexical norms*: that is, what words are used and understood, at what ages. Lexical norms can be estimated in different ways. One way is to take a corpus of transcribed children's language, extract word frequencies, and group the frequencies by age. For this to be possible, there must exist a corpus of children's language that is both comprehensive and machine readable. Such a corpus is CHILDES (MacWhinney, 2000), a freely available corpus of children's language containing transcriptions in over twenty languages. The English part alone comprises more than 5,000 transcriptions, most tagged with the age and gender of the children. Tools exist for exploring CHILDES, but none of them facilitate the calculation and visualization of word frequencies.

This report describes the implementation of ChildFreq, a computer program that makes exploring word frequencies in CHILDES easy. Frequencies can be ordered by

[1] http://childes.psy.cmu.edu/

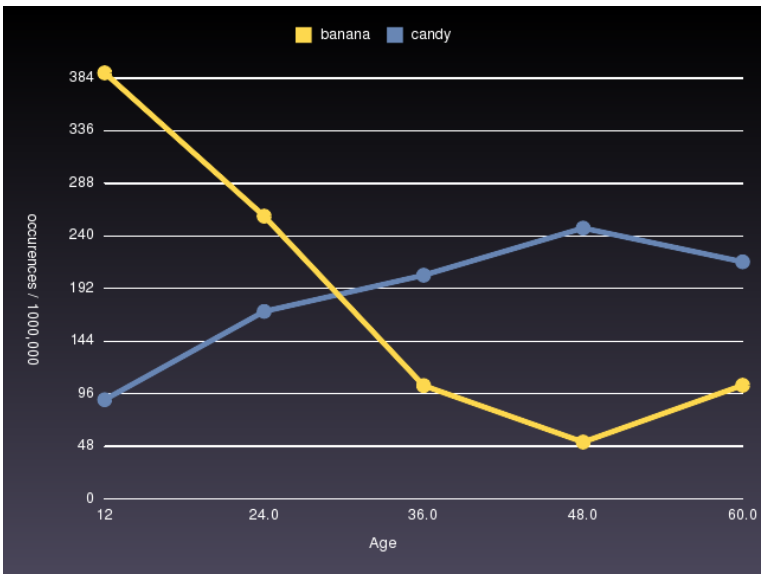

Figure 1: The line chart produced by ChildFreq given the query "`banana, candy`".

age or *mean length of utterance* (MLU) and it is also possible to split the data by the children's gender. ChildFreq is available online via any modern browser. It has a simple, intuitive interface and presents query results both as line charts and as more detailed tables. See Figure1 for an example of such a line chart .

The CLEX project (Jørgensen et al., 2009), another online tool for assessing lexical norms, takes a different approach. Where ChildFreq derives word frequencies from a standard corpus, CLEX uses information collected from surveys of parents asked to report what words, from a predefined set, their children can produce or understand. Consequently CLEX differs from ChildFreq in a number of important ways:

- Given a query word, while ChildFreq reports the number of word tokens per million words, CLEX reports the percentage of children who understand that word.
- CLEX can only be queried on words that were included in the parental survey: between 396 and 680 words, depending on the age of the children. ChildFreq can be queried on any word found in CHILDES.
- CLEX has data for children from eight to thirty months. ChildFreq has data for children from six to 96 months.

In sum: The two tools work in different ways on very different data (survey results versus corpus) and output different information. ChildFreq is not a substitute for but a complement to CLEX.

The next section describes in more detail the CHILDES database and what parts of it ChildFreq uses. Section 3 describes the implementation and outlines possibilities for further development. Section 4 describes ChildFreq in use. For those who want to get up and running quickly, a number of sample queries can be found in Section 4.1.

2 The CHILDES Database

The Child Language Data Exchange System (CHILDES; MacWhinney, 2000) is a corpus of child language data. It consists of transcriptions in over twenty languages though the greater number are in English. CHILDES is one component of the larger TalkBank corpus and shares some features common to all components, such as using the CLAN transcription tool and the CHAT transcription format.

A CHILDES transcription contains much information that ChildFreq does not use: phonetic transcription, excluded words, part-of-speech tagging, etc. ChildFreq

uses only the words (utterances) and the age, sex, and MLU of the child. The MLU is calculated from the sum of the number of morphemes uttered divided by the total number of statements. Spelling is standardized: e.g. if a child says "doggie", it is encoded as "dog". Transcriptions in CHILDES are a mix of words (utterances) and various forms of meta-information as mentioned above. For ChildFreq, CHILDES is preprocessed to remove everything but the children's words. So along with the meta-information, utterances from parents and researchers are also removed. A child is taken to be anyone identified as twelve years of age or younger or labeled as "CHI" or some variation of "child".

3 Implementation

ChildFreq consists of two parts: a back end that communicates with the database and makes the frequency calculations, and a front end that accepts queries and displays results.

The front end translates queries into the form:

```
SEARCH_STRING: <text>
SPLIT_SEXES: 0 | 1
X_AXIS: ''age'' | ''mlu''
AGE_GROUP: <number>
LOWER_AGE: <number>
UPPER_AGE: <number>
MLU_GROUP: <number>
LOWER_MLU: <number>
UPPER_MLU: <number>
```

`SEARCH_STRING` is the word or word pattern(s) to count. The allowable syntax is given is Section 4. `SPLIT_SEXES` is a boolean value that determines whether transcriptions from males and females should be counted separately, while `X_AXIS` determines whether the frequencies should be presented as a function of age or MLU. `AGE_GROUP` and `MLU_GROUP` specify how large the group size should be in the resulting graph. `LOWER_AGE`, `UPPER_AGE`, `LOWER_MLU`, and `UPPER_MLU` set the upper and lower boundaries on the transcriptions to be counted.

Given some query, the back-end begins by discarding those transcriptions not tagged with (depending on the value of `X_AXIS`) age or MLU, or outside the boundaries set by `LOWER_AGE`/`UPPER_AGE` or `LOWER_MLU`/`UPPER_MLU`. For each remaining transcription, the number of words matching `SEARCH_STRING` is calculated, then divided by the total number of words in the transcription, and multiplied by 1,000,000 to yield a measure of *word tokens per million*. These values are then grouped, averaged according to `AGE_GROUP` or `MLU_GROUP`, and displayed.

The front end is a web-interface, making ChildFreq available via any modern browser on any computer with an internet connection. To make a query, the user fills in a form (shown in Figure 2) that is accompanied by a short usage description. The results are returned as a line chart (as in Figure 1), and a table that can be copied and pasted into e.g. a spreadsheet program.

3.1 Future Development

ChildFreq is an ongoing project that will continue to be actively maintained and developed. As the CHILDES database continues to grow, the ChildFreq database will need to be updated. Plans are to do so at least annually. The ChildFreq homepage encourages users to send ideas for improvements and new features to the maintainer.

3.2 Implementation Language Details

Both the front and back end, are implemented in Ruby (`http://www.ruby-lang.org`). The pre-processed CHILDES transcriptions are stored in a database using YAML (`http://yaml.org`), a human readable data serialization format. The front end (web

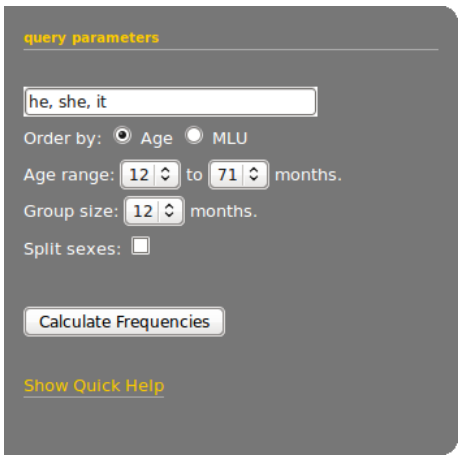

Figure 2: The search form.

interface) is built using the web framework Ruby on Rails (`http://rubyonrails.org`), and the charts are generated by the Gruff library (`http://gruff.rubyforge.org`). ChildFreq's source code is licensed under the open source MIT license (`http://www.opensource.org/licenses/mit-license.php`) and available at `http://childfreq.sumsar.net/source/childfreq_source.zip`.

4 Usage

Constructing queries in ChildFreq is meant to be easy, and the quick help accessible from the home page should be everything a user needs to get up and running. This section will describe the different search fields and what values they accept. See Figure 2.

The search bar

The search bar accepts one or more words. Allowable characters a–z, A–Z, and space. Words separated by spaces are treated as a single string: e.g., "I like" or "dog food". No distinction is made between upper- and lower-case letters; internally, all are treated as lower-case. Disallowed symbols will generate a warning.

Certain symbols have a special meaning: namely, "|" (*vertical bar*), "*" (*asterisk*), "," (*comma*) and "?" (*question mark*). A comma separates words that should be counted separately. A vertical bar separates words that should be grouped and counted together. The asterisk is a wild card that matches any number of characters (i.e., zero or more). A sample search string containing these symbols is:

> "dino*, dog|cat"

This would yield a two-line graph, with one line showing frequencies of words beginning with "dino", and the other showing frequencies of "dog" and "cat". The question mark can only be used on its own. The search string "?" counts the number of questions asked by the appropriate subset of children in CHILDES.

Order by

This can be set to either "Age" or "MLU" and determines whether word frequencies are ordered according to the age of the children or the MLU.

Age/MLU range

Depending on the value of "Order by", this allows the user to specify either the age range (in months) or the MLU range to include. This is useful when e.g. a user wants

to exclude the transcriptions of older children, of which there are quite few in the database.

Group size

Likewise dependent on the value of "Order by", "Group size" specifies the size of each data group, either in months or in units of MLU as appropriate. So for example, if ordering by age, a group size of 12 would yield a graph incremented in units of 12; the data point labeled "24" would include children from 24 to 36 months.

Split sexes

This splits the data by the gender of the children and results in a graph with twice as many lines as search terms. Any transcriptions not tagged with the gender of the child will be discarded in the search.

4.1 Examples of Usage

Here are a number of sample search queries showing what can be done with ChildFreq. Unless otherwise stated, the following values are assumed: `''Order by'' = ''Age''`, `''Age range'' = 12--71`, `''Group size'' = 12`, `''Split sexes'' = ''off''`.

- `dog, cat`
  Counts all occurrences of "cat" and "dog" separately, resulting in a line chart with two lines: one for "dog" and one for "cat".
- `dog | cat, dino*`
  Sums occurrences of "cat" and "dog", and separately counts all words beginning with "dino" (e.g., dino, dinosaur, dinosauria).
- `father|dad*, mother|mom*`
  Separately sums occurrences of "father", "dad", "dads", "daddy", etc.; and "mother", "mom", "moms", "mommy", etc.
- `pink, blue`
  `Split sexes = on`
  Separately counts occurrences of "pink" and "blue", and then separates the results by gender.
- `he|his|him, her|hers|she`
  `Split sexes = on`
  Separately sums occurrences of "he", "his", and "him"; and "her", "hers", and "she". The results are then separated by gender. Results in a line chart with four lines.
- `?`
  `Split sexes = on`
  Separately counts the number of questions asked by males and females.

5 Acknowledgment

ChildFreq was built as part of the VAAG project (`http://www.zas.gwz-berlin.de/research/projects/vaag/`) funded by the European Science Foundation, and the CCL project (`http://project.ht.lu.se/ccl/`) funded by the Swedish Research Council.

References

Golinkoff, R. M. and Hirsh-Pasek, K. (2008). How toddlers begin to learn verbs. *Trends Cogn. Sci. (Regul. Ed.)*, 12(10):397–403.

Goodman, J. C., Dale, P. S., and Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515–31.

Jørgensen, R., Dale, P., Bleses, D., and Fenson, L. (2009). Clex: A cross-linguistic lexical norms database. *Journal of child language*, pages 1–10.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates.