

Doing corpus phonetics using data on the web

고연숙

조선대학교

eonsukko@chosun.ac.kr

Data for studying mother-child interaction: Introducing the CHILDES database

<https://childes.talkbank.org>

Overview of the CHILDES system

- CHILDES = **C**hild **L**anguage **D**ata **E**xchange **S**ystem
- Basic components of CHILDES
 - A database of child language transcripts
 - CHAT = a system of **C**odes for the **H**uman **A**nalysis of **T**ranscripts of child speech
 - CLAN = a collection of **C**hild **L**anguage **A**nalysis programs

Database

- Variety of languages
- Typically developing populations
- Atypically developing populations
 - Down syndrome
 - Autism
 - Specific language impairment
 - Brain lesions

CHAT: Codes for the human analysis of transcripts

@Begin

@Participants: ROS Ross Child, BRI Brian Father

*ROS: why isn't Mommy coming?

%com: Mother usually picks Ross up around 4 pm.

*BRI: don't worry.

*BRI: she'll be here soon.

*ROS: good (be)cause I'm xxx hungry.

*BRI: what would you like to eat?

@End

CLAN: Child Language Analysis Program

- **FREQ:** count items
- **KWAL:** search for individual keywords in context
- **MLU:** mean length of utterance

Download the program at
<http://childes.talkbank.org>

Command syntax

<u>mlu</u>	<u>+t*CHI</u>	<u>adam01.cha</u>
command	option	filename

- **mlu**: the name of the program you wish you execute
- **+t*CHI**: a list of options you select for the program
- **adam01.cha**: the name of the transcripts you wish to analyze

options

- +t: speaker code
 - freq +t*CHI adam01.cha
 - freq -t*CHI adam01.cha
- +s: focus analysis on particular strings
 - freq +s"who" adam01.cha
 - freq -s"who" adam01.cha
 - kwal +s"who" adam01.cha

- *: wild card
 - freq +t*CHI +s"toy" adam01.cha
 - freq +t*CHI +s"toy*" adam01.cha
 - freq +t*CHI +s"toy*" adam0*.cha
- +u: perform a combined analysis
 - freq +t*CHI +s"toy*" +u adam0*.cha
- w: provides a window for the context
 - kwal +t*MOT -w2 +w2 +sno adam01.cha
- @: include file
 - kwal +s@whwords adam01.cha

Multiple keywords in a file (whwords.txt)

- Open a text editor
- The contents of the file:
 - who
 - where
 - why
 - where
 - what
- Save it in a .txt format





```
freq +t*CHI +s@whwords.txt *.cha
```

- The above command will search for the frequency of the words included in the super.txt file.
- **+t*CHI**: analyze the child's tier only
- **+s@whwords.txt**: include the words from the file whwords.txt
- ***.cha**: analyze all files ending in .cha
- **+u**: adding this option will merge all the input files into a combined analysis

Sound symbolism

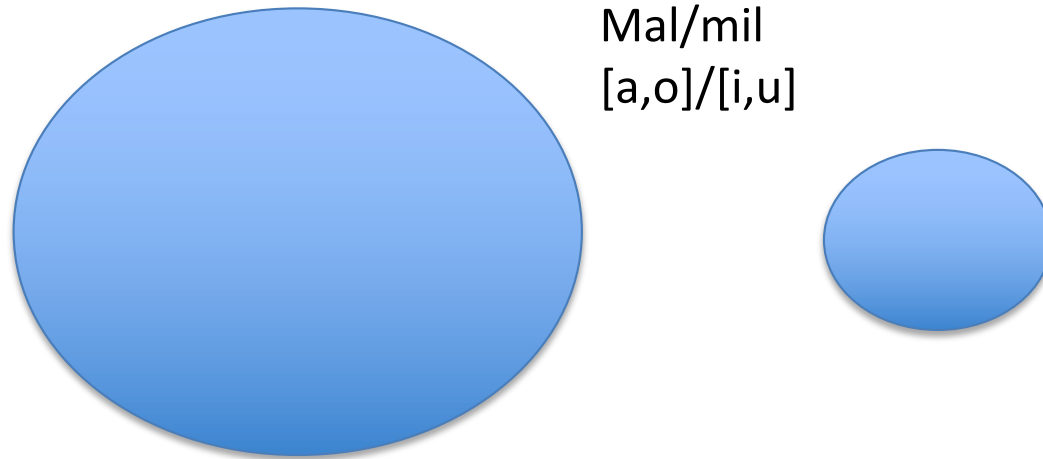
Three types of signs

Charles Saunders Peirce (1838-1914)

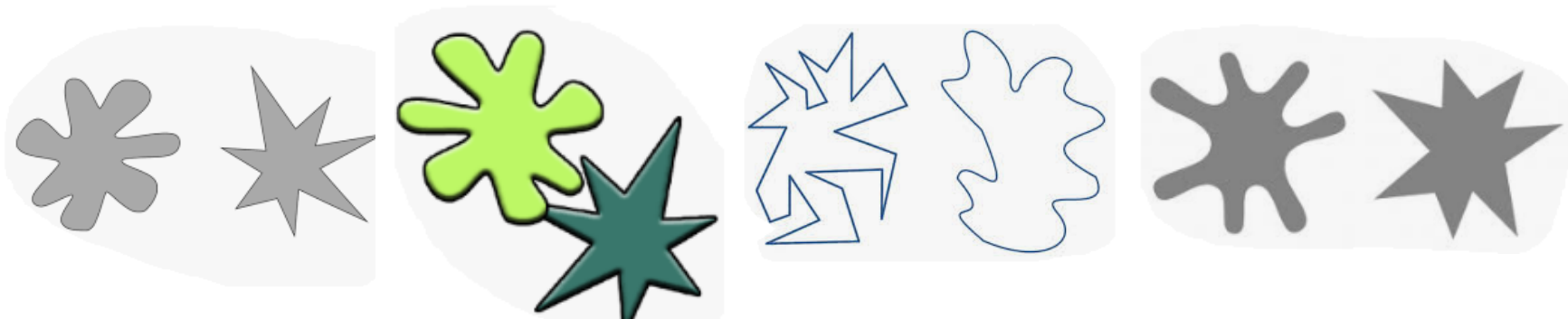
- **Icons**: signs that *resemble* the things they represent
 - e.g., mirror images, photos, paintings, maps  
- **Indices**: signs that are spatially, temporally, or causally related to what they represent
 - e.g., smoke as a sign of fire; arrows indicating directions  
- **Symbols**: the relationship between the form and meaning is arbitrary.

Sound symbolism

- Some words are classified as 'iconic', rather than symbolic.
 - The sound of a word represents its lexical meaning.



Takete/baluba (maluma), kiki/bouba (Köhler1929)



✓ **Sound symbolism bootstrapping hypothesis** (Imai and Kita, 2014)

“sound symbolism provides a scaffolding mechanism for children in various stages of language development.”

- Infants are sensitive to sound symbolism, due to naturally occurring multi-modal mapping (e.g., sound + vision).

- Sound symbolism helps infants realize that speech sounds refer to entities in the world.

- We can hypothesize that words with sound symbolism (e.g. woof, or hu:ge) might help infants to learn vocabulary because of the **close resemblance in form and meaning** in the expressions.
- Words that resemble the sound of manner of certain meaning conveyed by words are called **mimetics** or **onomatopoeia**.
- If the hypothesis above were true, we might expect parents to use sound symbolism **more frequently to younger children** than older. Let's investigate this question.

- Find words with sound symbolism in parents' speech. Using the "freq" command, generate a list of all words that the mother produced.

freq +t*MOT *.cha > freq.txt

This command will give you a list of words in an alphabetical order.

- If you want to view the list in the order of frequency so that the most frequently used words appear on the top, run the following command.

freq +t*MOT +o *.cha > freq_o.txt

What are the most frequent 5 mimetic words in the freq_o.txt output?

- We would like to know how often sound symbolism is used in mothers' speech as a function of child's age. For this task, make a file called "mimetic.txt" which contains a list of words that you will be searching for in each file.
- If you run the following command, it will extract all the mimetic words contained in the mimetic.txt file from Naima's mother's tier.

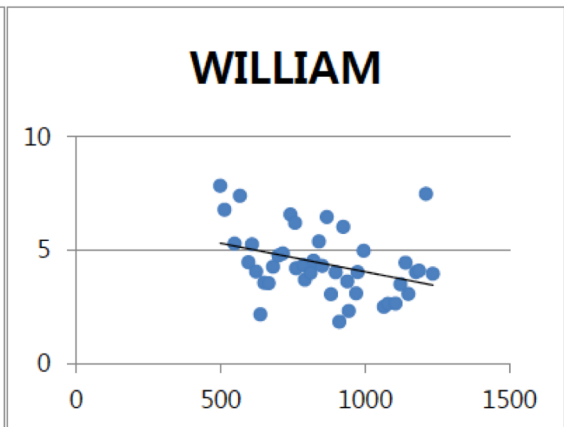
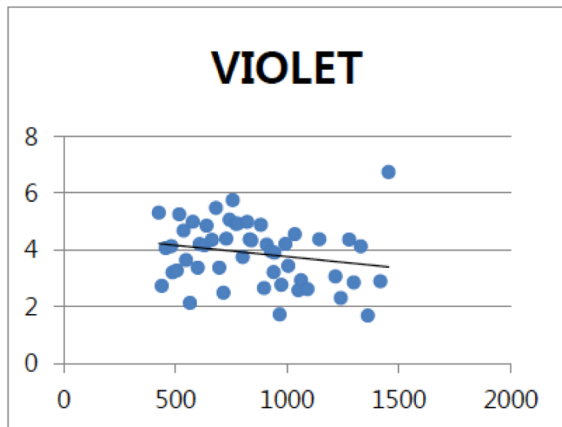
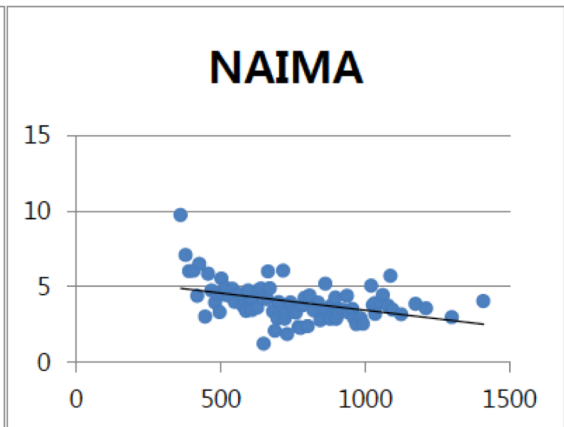
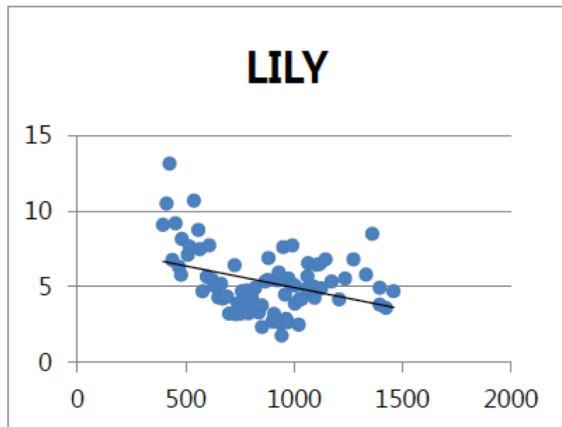
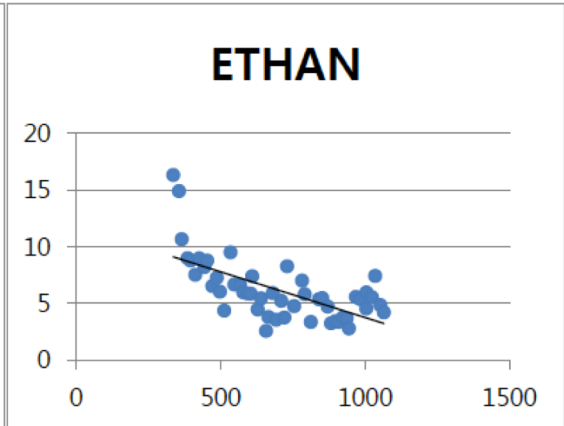
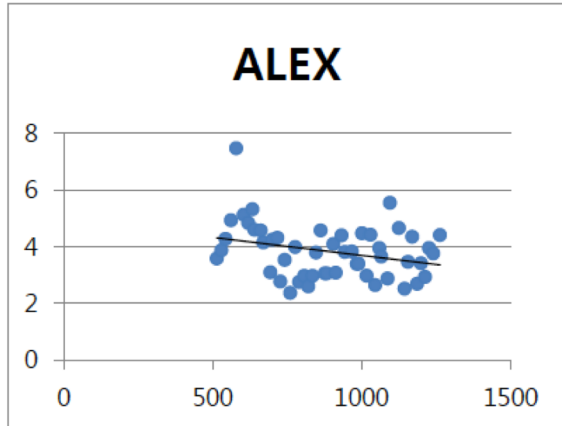
freq +t*MOT +u +s@mimetic.txt *.cha, etc.

- Now add "+3" option to generate an excel file that contains the same information.

freq +t*MOT +d3 +s@mimetic.txt *.cha, etc.

Report the number of mimetic words in each file for each mother.

- In order to account for differences in the amount of speech across parents, we need to determine **how many words each parents produced**. The MLU program generates this information.
- **mlu +t*MOT -t%mor +d *.cha**
- Report the number of total words (not morphemes) for each of the six mothers.
- Plot a graph where x-axis has information on each child and y-axis on the usage of the mimetic words per 100 words.



CLAN and Praat

- CLAN 에서 kwal 명령어를 이용하여 분석하고자 하는 단어가 포함된 문장을 뽑는다.
 - kwal +t*MOT +s"book" *.cha
 - kwal +x=5w +t*MOT 011116.cha
- Praat 을 구동한다.
- Send to sound analyzer 명령어를 이용하여 Praat 의 object 창으로 해당 음성 구간을 이동한다.
- ObjectstoFiles 스크립트를 이용하여 음성 파일로 저장한다.
- Grid-maker 스크립트를 이용하여 코딩한다.
- Analyze_tier 스크립트를 이용하여 음장, 피치, 포먼트 등의 값을 추출한다.

Kwal 을 이용해 문장 추출하기

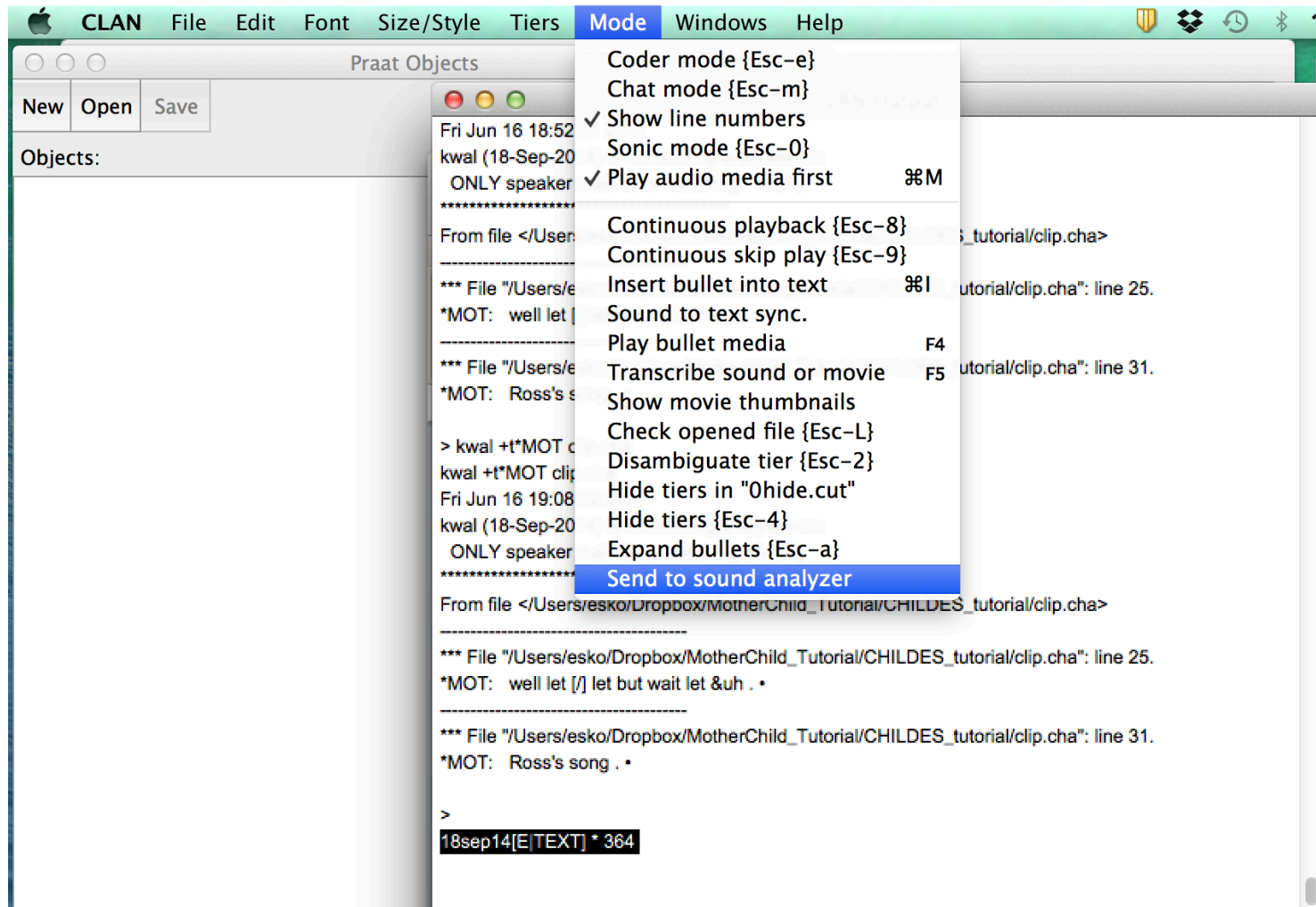
The image shows two overlapping windows from a CLAN software interface. The left window, titled "CLAN Output", displays the results of a text analysis. It shows two separate analysis runs. The first run, dated "Fri Jun 16 18:52:31 2017", identifies two sentences: "*MOT: well let [/] let but wait let &uh . ." and "*MOT: Ross's song . .". The second run, dated "Fri Jun 16 19:08:22 2017", identifies the same two sentences. At the bottom of the CLAN Output window, a terminal prompt ">" is followed by the command "18sep14[E|TEXT] * 365". The right window, titled "Commands", is a configuration panel for a command. It lists several commands with their corresponding paths: "working" at "/Users//D.../MotherChild_Tutorial/CHILDES_tutorial", "output" at "/Users//Dropbox/연구재단/Korea-America/VOT_0626", "lib" at "/Applications/CLAN/lib/", and "mor lib" at "/Applications/CLAN/lib/". Below the list is a "Progs" dropdown menu with a question mark icon. At the bottom of the Commands window, there are "Recall" and "Run" buttons, with "18sep14" displayed between them.

```
Fri Jun 16 18:52:31 2017
kwal (18-Sep-2014) is conducting analyses on:
  ONLY speaker main tiers matching: *MOT;
*****
From file </Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha>
-----
*** File "/Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha": line 25.
*MOT: well let [/] let but wait let &uh . .
-----
*** File "/Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha": line 31.
*MOT: Ross's song . .

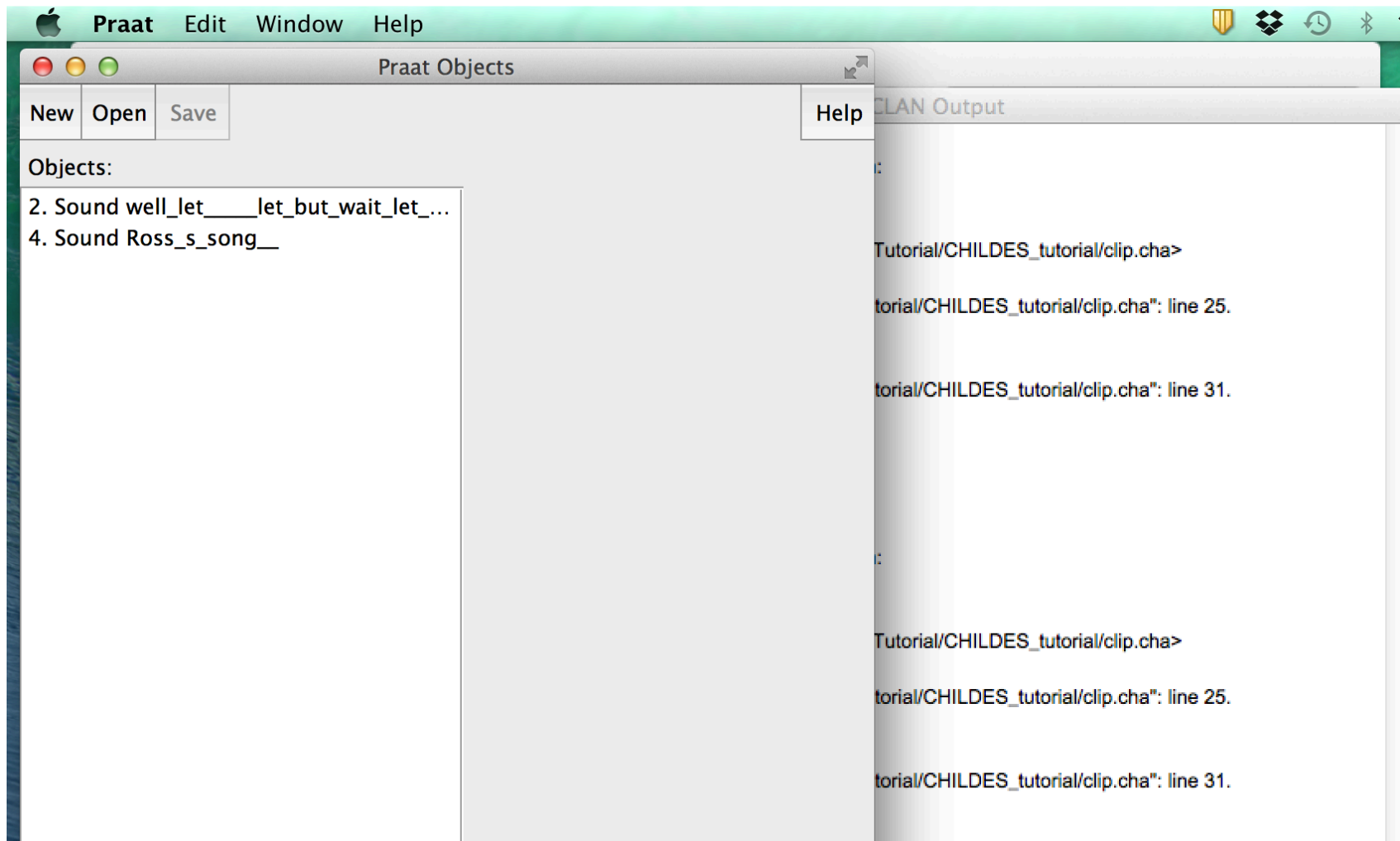
> kwal +t*MOT clip.cha
kwal +t*MOT clip.cha
Fri Jun 16 19:08:22 2017
kwal (18-Sep-2014) is conducting analyses on:
  ONLY speaker main tiers matching: *MOT;
*****
From file </Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha>
-----
*** File "/Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha": line 25.
*MOT: well let [/] let but wait let &uh . .
-----
*** File "/Users/esko/Dropbox/MotherChild_Tutorial/CHILDES_tutorial/clip.cha": line 31.
*MOT: Ross's song . .

>
18sep14[E|TEXT] * 365
```

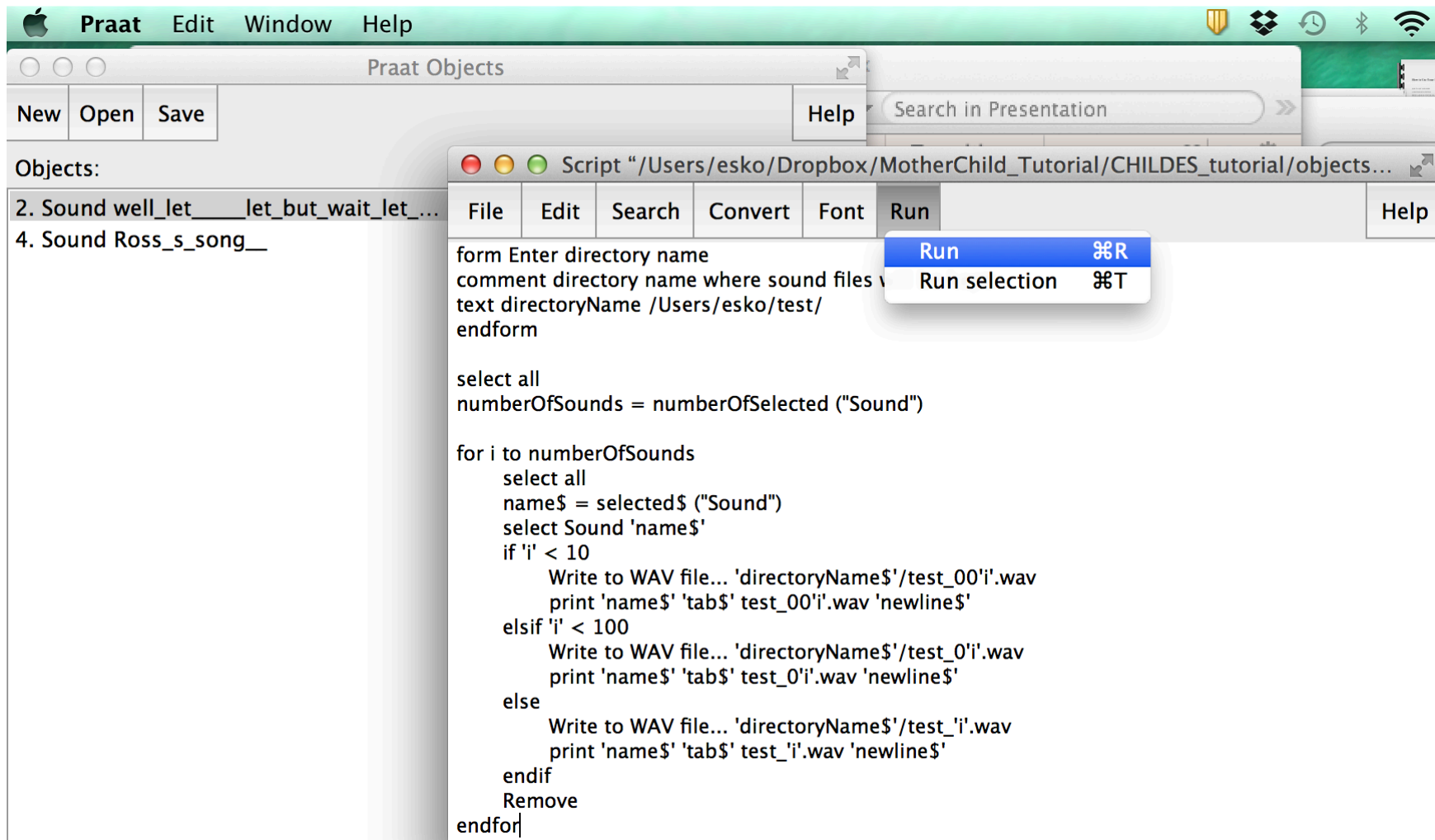
Send to sound analyzer



Sound objects created in praat



Save the sound objects as files



Annotate the target vowels

The screenshot displays the Praat software interface. The main window shows two audio waveforms and a spectrogram. The spectrogram has a red horizontal line at 1175 Hz. A blue vertical bar highlights a segment of the audio, with time markers 0.749689, 0.073, and 0.823011. A text grid below the spectrogram shows a yellow box with the letter 'v' under the highlighted segment. A dialog box is open in the foreground with the text "Annotate tiers, then press continue..." and "Continue" button highlighted.

Objects:

- 5. Strings list
- 6. Sound_test
- 7. TextGrid_test

Script "..."

```
## Praat script by Ke...
## Parts inspired by ...
## Below: user provi...
## you will probably ...
## filename minus th...

form Select directory
sentence Directo...
sentence Filenar...
sentence Extens...
sentence Tier(s)...
endform

Create Strings as file...
file_count = Get num...

## Loop through file...

for k from 1 to file_c...
select Strings list
current$ = Get st...
Read from file... '
short$ = selected
```

7. TextGrid_test_001

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses

0.2996
-0.8398
0.2996
-0.8398
5000 Hz
1175 Hz
0 Hz

0.749689 0.073 0.823011

Pause: stop or continue

Annotate tiers, then press continue...

Stop Continue

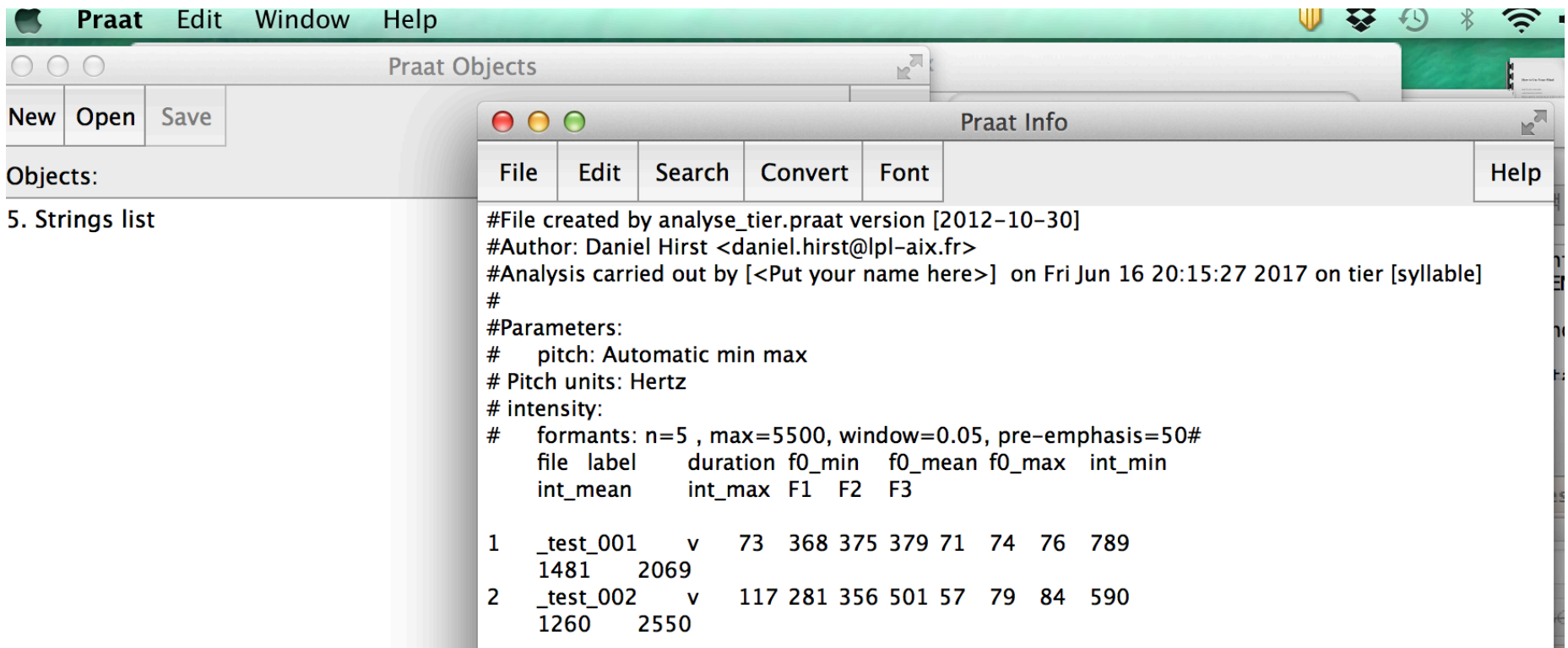
Filename

0 0.749689 0.073 0.28798

Visible part 1.110998 seconds

Total duration 1.110998 seconds

Extract the results



The screenshot shows the Praat software interface. The main window displays "5. Strings list" under the "Objects:" section. A "Praat Info" window is open, showing the following text:

```
#File created by analyse_tier.praat version [2012-10-30]
#Author: Daniel Hirst <daniel.hirst@lpl-aix.fr>
#Analysis carried out by [<Put your name here>] on Fri Jun 16 20:15:27 2017 on tier [syllable]
#
#Parameters:
#  pitch: Automatic min max
# Pitch units: Hertz
# intensity:
#  formants: n=5 , max=5500, window=0.05, pre-emphasis=50#
file label  duration f0_min  f0_mean f0_max  int_min
int_mean   int_max  F1  F2  F3
1  _test_001  v    73  368 375 379 71  74  76  789
   1481  2069
2  _test_002  v    117 281 356 501 57  79  84  590
   1260  2550
```

- For more information, please refer to the CLAN manual at talkbank.org