This is a file of the commands that were hidden in CLAN on March 5, 2019. They are still there and operative, but they are no longer listed or documented.

1.1 COMPOUND

This program changes pairs of words to compounds, to guarantee more uniformity in morphological and lexical analysis. It requires that the user create a file of potential compound words in a format with each compound on a separate line, as in this example.

```
night+night
Chatty+baby
oh+boy
```

Whenever the program finds “night night” in the text, whether it be written as “night+night”, “night night” or “night-night,” it will be changed to “night+night”.

1.2 CMDI

This program runs over the CHILDES and TalkBank databases to produce a complete metadata inventory in CMDI format.

1.3 COMBINE

This program combines multiple files that coded participants separately into a single file based on the time codes.

1.4 DATACLEAN

DATACLEAN is used to rearrange and modify old style header tiers and line identifiers.

1. If @Languages tier is found, it is moved to the position right after @Begin.
1. If @Participants tier is found, it is moved to the position right after @Languages.
2. If the tier name has a space character after ':', then it is replaced with tab. If the tier name doesn't have a following tab, then it is added. If there is any character after the tab following a speaker name, such as another tab or space, then it is removed.
3. Tabs in the middle of tiers are replaced with spaces.
4. If utterance delimiters such as +... are not separated from the previous word with a space, then a space is inserted.
5. if [...] is not preceded or followed by space, then space is added.
6. Replaces #long with ###.
7. The string "..." is replaced with "+...".

1.5 FIXLANG

Changes the codes for languages to the three-letter ISO standard.

1.6 FIXMP3S

This program fixes errors in the time alignments of sound bullets for MP3 files in versions of

CLAN before 2006.

1.7 JOINITEMS

This program finds the items specified by the +s switch and joins them with the previous words. It is useful for reformatting Japanese particles as suffixes.

1.8 OLAC

This program goes through the various directories in CHILDES and TalkBank and creates an XML database that can be used by the OLAC (Online Language Archives Community) system to help researchers locate corpora relevant to their research interests.

1.9 ORT

ORT, if +c used, then this converts HKU style disambiguated pinyin with capital letters to CMU style lowercase pinyin on the main line. Without the +c switch, it is used to create a %ort line with Hanzi characters corresponding to the pinyin-style words found on main line. The choice of characters to be inserted is determined by entries in the lexicon files at the end of each word's line after the '%' character.

1.10 PHONFREQ

The PHONFREQ program tabulates all of the segments on the %pho line. For example, using PHONFREQ with no further options on modrep.cha will produce this output:

```
2 A    initial =   0, final =   1, other =   1
1 I    initial =   0, final =   1, other =   0
3 a    initial =   1, final =   1, other =   1
2 m    initial =   2, final =   0, other =   0
3 n    initial =   1, final =   1, other =   1
2 o    initial =   0, final =   2, other =   0
2 w    initial =   2, final =   0, other =   0
```

This output tells you that there were two occurrences of the segment /A/, once in final position and once in other or medial position.

If you create a file called alphabet file and place it in your working directory, you can further specify that certain digraphs should be treated as single segments. This is important if you need to look at diphthongs or other digraphs in UNIBET. In the strings in the alphabet file, the asterisk character can be used to indicate any single character. For example, the string *: would indicate any sound followed by a colon. If you have three instances of a:, three of e:, and three of o:, the output will list each of these three separately, rather than summing them together as nine instances of something followed by a colon. Because the asterisk is not used in either UNIBET or PHONASCII, it should never be necessary to specify a search for a literal asterisk in your alphabet file. A sample alphabet file for English is distributed with CLAN. PHONFREQ will warn you that it does not find an alphabet file. You can ignore this warning if you are convinced that you do not need a special alphabet file.

If you want to construct a complete substitution matrix for phonological analysis, you need to add a %mod line in your transcript to indicate the target phonology. Then you can run PHONFREQ twice, first on the %pho line and then on the %mod line. To run on the %mod line, you need to add the +t%mod switch.

If you want to specify a set of digraphs that should be treated as single phonemes or segments, you can put them in a file called alphabet.cut. Each combination should be entered by itself on a single line. PHONFREQ will look for the alphabet file in either the working directory or the library directory. If it finds no alphabet.cut file, each letter will be treated as a single segment. Within the alphabet file, you can also specify trigraphs that should override particular digraphs. In that case, the longer string that should override the shorter string should occur earlier in the alphabet file.

1.10.1 Unique Options

The best way to see a complete list of options for a command is to type the name of the command followed by a carriage return in the Commands window. For example, if you type just the word **phonfreq**, you will see a list of all available options. Many of these will be options shared with other programs. For information on these, the best approach is to go to the chapter 8 in this manual which describes all these shared options.

In addition, many of the programs have some unique options. PHONFREQ has the following unique options:

+b By default, PHONFREQ analyzes the %pho tier. If you want to analyze another tier, you can use the +b switch to specify the desired tier. Remember that you might still need to use the +t switch along with the +b switch as in this command:

```
phonfreq +b* +t*CHI modrep.cha
```

+d If you use this switch, the actual words that were matched will be written to the output. Each occurrence is written out.

+t You should use the +b switch to change the identity of the tier being analyzed. The +t switch is used to change the identity of the speaker being analyzed. For example, if you want to analyze the main lines for speaker CHI, you would use this command:

```
phonfreq +b* +t*CHI modrep.cha
```

PHONFREQ also uses several options that are shared with other commands. For a complete list of options for a command, type the name of the command followed by a carriage return in the Commands window. Information regarding the additional options shared across commands can be found in the chapter on Options.

The lexicon could be much smaller if more rules were written to handle derivational morphology. These would handle prefixes such as “non#” and derivational suffixes such as “-al.” The grammar still needs to be fine-tuned to catch common over-regularizations, although it will never be able to capture all possible morphological errors. Furthermore, attempts to capture over regularizations may introduce bogus analyses of good forms, such as “seed” = “*see-PAST.” Other areas for which more rules need to be written include diminutives, and words like “oh+my+goodness,” which should automatically be treated as communicators.

1.11 SPREADSHEET

This program rotates a table by turning columns into rows and vice versa, or it can merge rows with matching cells into one row.

1.12 USEDLEX

This program removes all the items in the lexicon files of a MOR grammar that are not being used in a given corpus. This is helpful for work with transcriptions that use a limited target vocabulary.

1.13 SUBTITLES

This program converts subtitle files into CHAT format.

1.14 SYNCODING

This program is not functional yet.

1.15 UNIQ

UNIQ is used to sort lexicon files into alphabetical order, while removing duplicates.