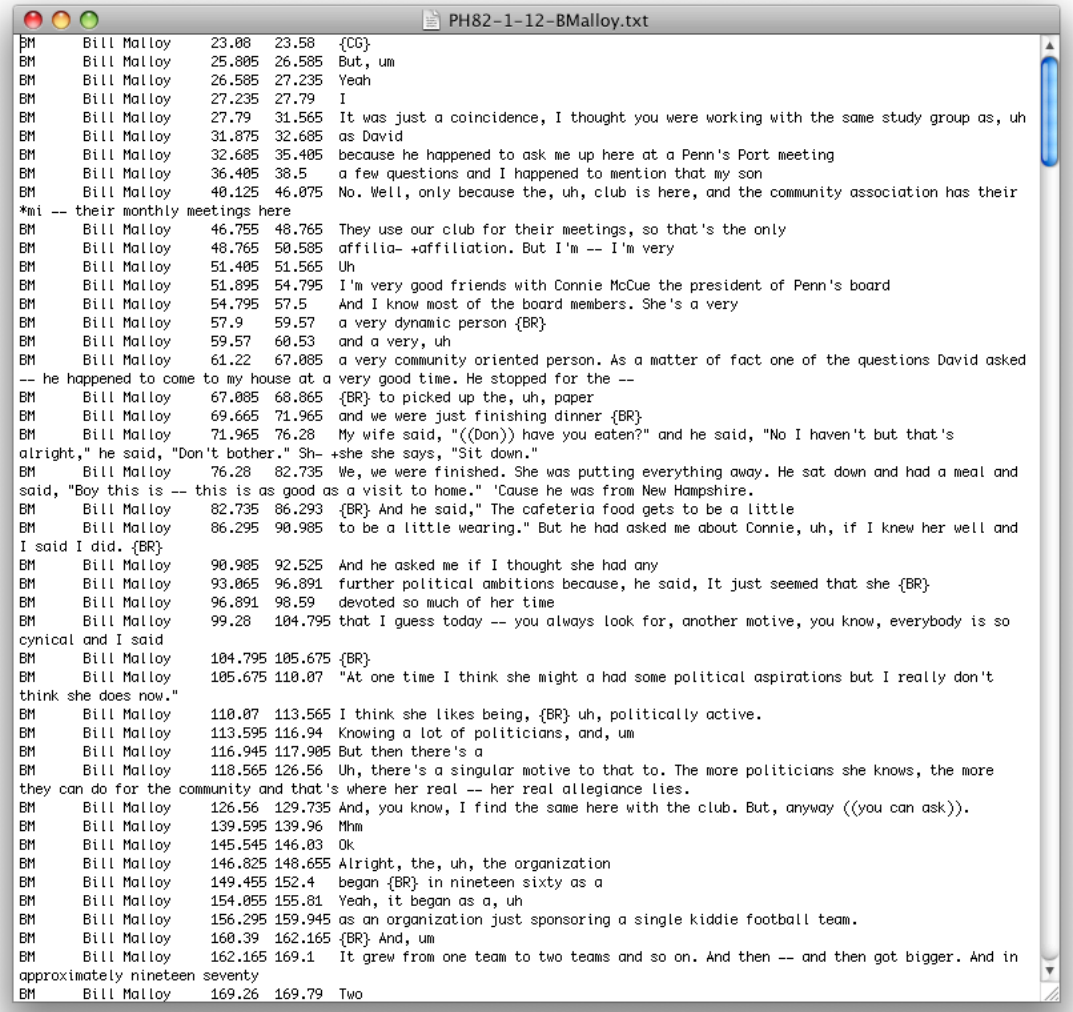
How to do forced alignment using *FAAValign*

1. Export a **transcript** from ELAN as **tab-delimited text file** (via *File > Export As > Tab-delimited Text…* – see ELAN introduction, Appendix A, for details on the format). This produces a **transcript file**, e.g. `PH82-1-12-BMalloy.txt`:



PH82-1-12-BMalloy.txt

```
BM      Bill Malloy     23.08   23.58   {CG}
BM      Bill Malloy     25.805  26.585  But, um
BM      Bill Malloy     26.585  27.235  Yeah
BM      Bill Malloy     27.235  27.79   I
BM      Bill Malloy     27.79   31.565  It was just a coincidence, I thought you were working with the same study group as, uh
BM      Bill Malloy     31.875  32.685  as David
BM      Bill Malloy     32.685  35.405  because he happened to ask me up here at a Penn's Port meeting
BM      Bill Malloy     36.405  38.5    a few questions and I happened to mention that my son
BM      Bill Malloy     40.125  46.075  No. Well, only because the, uh, club is here, and the community association has their
*mi -- their monthly meetings here
BM      Bill Malloy     46.755  48.765  They use our club for their meetings, so that's the only
BM      Bill Malloy     48.765  50.585  affilia- +affiliation. But I'm -- I'm very
BM      Bill Malloy     51.405  51.565  Uh
BM      Bill Malloy     51.895  54.795  I'm very good friends with Connie McCue the president of Penn's board
BM      Bill Malloy     54.795  57.5    And I know most of the board members. She's a very
BM      Bill Malloy     57.9    59.57   a very dynamic person {BR}
BM      Bill Malloy     59.57   60.53   and a very, uh
BM      Bill Malloy     61.22   67.085  a very community oriented person. As a matter of fact one of the questions David asked
-- he happened to come to my house at a very good time. He stopped for the --
BM      Bill Malloy     67.085  68.865  {BR} to picked up the, uh, paper
BM      Bill Malloy     69.665  71.965  and we were just finishing dinner {BR}
BM      Bill Malloy     71.965  76.28   My wife said, "((Don)) have you eaten?" and he said, "No I haven't but that's
alright," he said, "Don't bother." Sh- +she she says, "Sit down."
BM      Bill Malloy     76.28   82.735  We, we were finished. She was putting everything away. He sat down and had a meal and
said, "Boy this is -- this is as good as a visit to home." 'Cause he was from New Hampshire.
BM      Bill Malloy     82.735  86.293  {BR} And he said," The cafeteria food gets to be a little
BM      Bill Malloy     86.295  90.985  to be a little wearing." But he had asked me about Connie, uh, if I knew her well and
I said I did. {BR}
BM      Bill Malloy     90.985  92.525  And he asked me if I thought she had any
BM      Bill Malloy     93.065  96.891  further political ambitions because, he said, It just seemed that she {BR}
BM      Bill Malloy     96.891  98.59   devoted so much of her time
BM      Bill Malloy     99.28   104.795 that I guess today -- you always look for, another motive, you know, everybody is so
cynical and I said
BM      Bill Malloy     104.795 105.675 {BR}
BM      Bill Malloy     105.675 110.07  "At one time I think she might a had some political aspirations but I really don't
think she does now."
BM      Bill Malloy     110.07  113.565 I think she likes being, {BR} uh, politically active.
BM      Bill Malloy     113.595 116.94  Knowing a lot of politicians, and, um
BM      Bill Malloy     116.945 117.905 But then there's a
BM      Bill Malloy     118.565 126.56  Uh, there's a singular motive to that to. The more politicians she knows, the more
they can do for the community and that's where her real -- her real allegiance lies.
BM      Bill Malloy     126.56  129.735 And, you know, I find the same here with the club. But, anyway ((you can ask)).
BM      Bill Malloy     139.595 139.96  Mhm
BM      Bill Malloy     145.545 146.03  Ok
BM      Bill Malloy     146.825 148.655 Alright, the, uh, the organization
BM      Bill Malloy     149.455 152.4   began {BR} in nineteen sixty as a
BM      Bill Malloy     154.055 155.81  Yeah, it began as a, uh
BM      Bill Malloy     156.295 159.945 as an organization just sponsoring a single kiddie football team.
BM      Bill Malloy     160.39  162.165 {BR} And, um
BM      Bill Malloy     162.165 169.1   It grew from one team to two teams and so on. And then -- and then got bigger. And in
approximately nineteen seventy
BM      Bill Malloy     169.26  169.79  Two
```

Note: It is important that the input text file has this exact format. If there are more than five fields per line (one each for name of the tier, name of the participant, beginning and end of the annotation unit in seconds, and the transcription text) – for example, if you forget to uncheck the "duration" check box in ELAN – *FAAValign.py* will not work.

2. Move the sound file (e.g. `PH82-1-12-BMalloy.wav`) and the transcript file (e.g. `PH82-1-12-BMalloy.txt`) into the forced alignment directory.

3. Open a new terminal window. Use the `cd` command to go to the forced alignment directory, e.g.

```
cd /Users/Shared/Forced_Alignment_Toolkit/
```

4. Run *FAAValign.py* with the **"dictionary check"** (***–check***) option:

```
python FAAValign.py -v -c unknownBM.txt PH82-1-12-BMalloy.txt[1]
```

This option performs a dictionary lookup for all words in the input transcription text to check whether a given word has an entry in the CMU pronouncing dictionary. All words in the input transcription text for which no entry is found in the dictionary, as well as all truncated words, are written to the file that is specified after `-c` (in the example above, `unknownBM.txt`). There are no restrictions on the name of the unknown words file; in the example above, the file name consists of "unknown" followed by the main speaker's initials.

The **–v ("verbose")** option produces verbose output. This is useful to know what's going on. (See the appendix for examples.)

This will produce a file with a **list of unknown words** `unknownBM.txt` with four columns:

A. unknown or truncated word
B. phonemic transcription
C. "clue word" (if in transcript)[2]
D. text of the annotation unit containing the unknown or truncated word

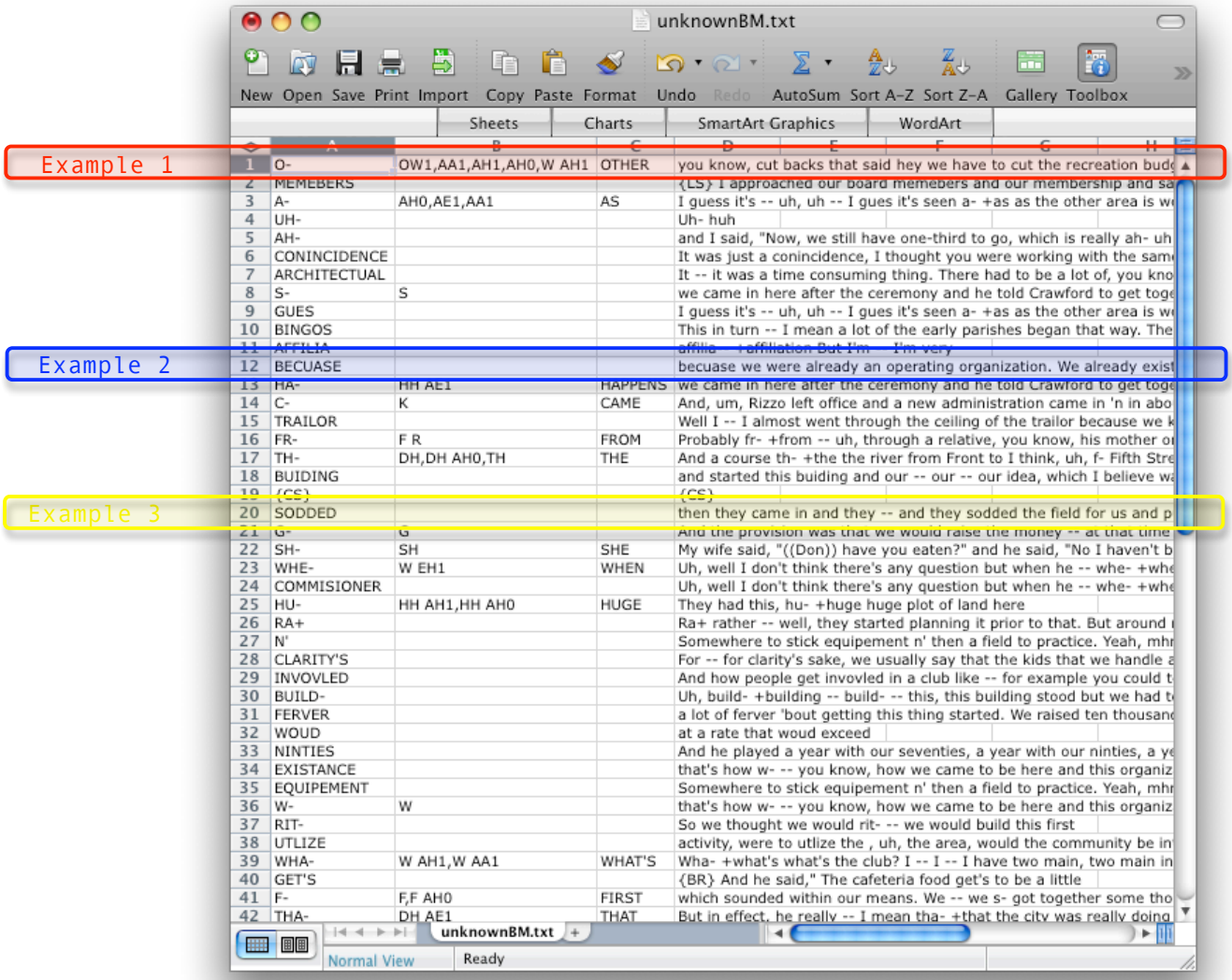
---

[1] Since you're only checking the transcript in this first step, you only need to specify the name of the transcript file. If the sound file and the transcript file have identical names except for the extensions, then you can also use the name of the sound file instead. (This was the setup in earlier versions of FAAValign.)

[2] A "**clue word**" is a word beginning with a **plus sign** which has been inserted by the transcriber after a truncated word if the transcriber is reasonably sure that this is the word the speaker intended to say.

Its purpose is to help the person aligning the sound file in determining what should be the correct phonemic transcription for the truncated word without having to go back to the original ELAN file and listen to the annotation unit in question. The clue word can be thought of as a sort of editorial comment inserted by the transcriber. It will be removed from the transcript text by *FAAValign* prior to alignment.

In the example screenshot below, line 25 contains both a truncated word and a clue word following immediately afterwards in the annotation unit They had this **hu- +huge** huge plot of land here. Column B shows that the suggested transcriptions for the truncated word "hu-" are "HH AH1" and "HH AH0", both of which have incorrect vowels (the strut vowel and schwa, respectively). Therefore, the transcription in column B should be replaced with the correct transcription "HH Y UW1".



5. Open `unknownBM.txt` in **Excel** (or any other spreadsheet application). Go through the entries one by one:

   a. If the entry is a **truncated word** and a **transcription** is **suggested** in column B of the spread sheet, check if this transcription is correct for the word in question. (A "clue word" entry in the column C might help in determining this.) → **Example 1**

      i. If the suggested transcription is **correct**, nothing needs to be done about it, so **delete the line**.

      ii. If it is **not** correct, or if there is no suggested transcription in the second column, enter the correct **transcription** in Arpabet format into the second column. You can enter several

transcription alternatives separated by commas. Make sure to enter stress digits for all vowels. For truncated words, it can help to copy the full form from the CMU pronouncing dictionary, if present, and truncate it down to wherever the truncation occurs in the entry.

b. If the entry is due to a **spelling mistake** in the original transcription → **Example 2**, go back to ELAN and **change the transcription**. (Using ELAN's "Find" option will help locate the annotation unit in question.) Save the corrected `.eaf` file and **delete the line** in question in `unknownBM.txt`.

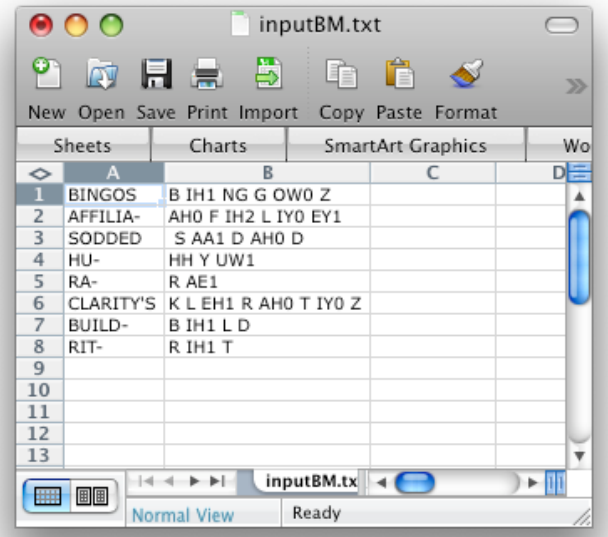
c. If the entry is simply **unknown** → **Example 3**, provide the **transcription** in Arpabet format. Again, make sure to enter stress digits for all vowels. If the word is unfamiliar (e.g. a proper name), it can help to go back to the original ELAN file, search for the word in question, and listen to the speaker's pronunciation of it for a couple of times.

6. **Delete** the **third** ("clue word") and **fourth** ("line") **columns** in `unknownBM.txt`.

7. **Save** the remaining file as a tab-delimited text file under **another name** (suggestion: `inputBM.txt`). This will be your dictionary input file. Put the input file into the forced alignment folder.

8. **Close Excel**.

9. Export a new version of the updated `.eaf` file from ELAN to produce a **new, updated input transcription file**. Put the new transcription file into the forced alignment directory.



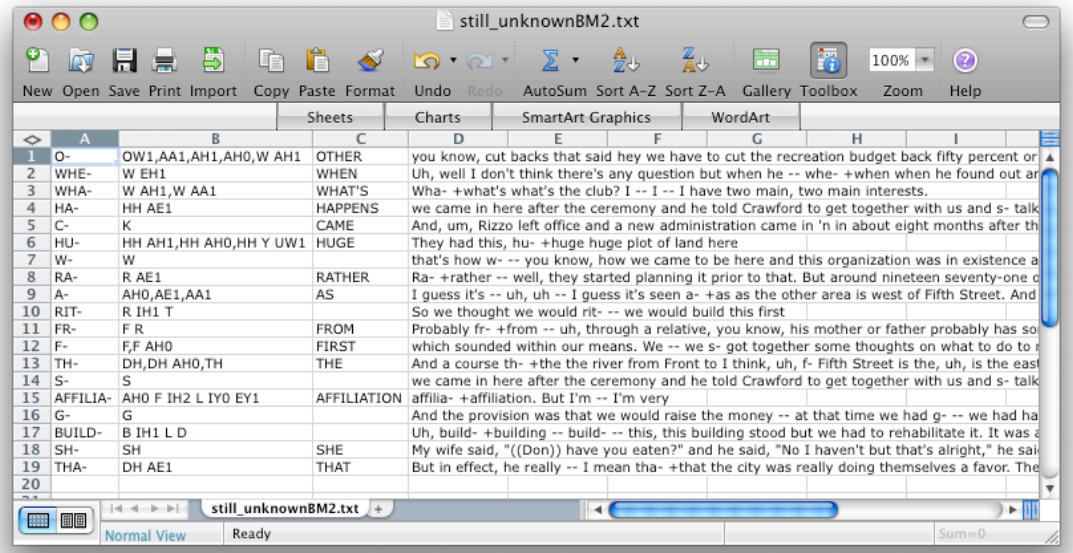
10. Run a second dictionary check to make sure that you really have supplied the transcriptions for all unknown words by running *FAAValign.py* with the **"import dictionary entries" (*–import*)** option for `inputBM.txt` and the updated input transcription file:

```
python FAAValign.py -v -i inputBM.txt -c still_unknownBM.txt PH82-1-12-
       BMalloy.txt
```

The "import" option will cause all entries in `inputBM.txt` to be **added to the CMU pronouncing dictionary** prior to alignment. ~~The updated version of~~

~~the dictionary will be written to file.~~[3] The transcriptions from the input file will also be added to a file `added_dict_entries.txt`, where they can later be edited manually and merged with the main dictionary.

To make sure that all unknown words are accounted for, check `still_unknownBM.txt`. The file should now only contain truncated words and their suggested transcriptions.



	A	B	C	D
1	O-	OW1,AA1,AH1,AH0,W AH1	OTHER	you know, cut backs that said hey we have to cut the recreation budget back fifty percent or
2	WHE-	W EH1	WHEN	Uh, well I don't think there's any question but when he -- whe- +when when he found out ar
3	WHA-	W AH1,W AA1	WHAT'S	Wha- +what's what's the club? I -- I -- I have two main, two main interests.
4	HA-	HH AE1	HAPPENS	we came in here after the ceremony and he told Crawford to get together with us and s- talk
5	C-	K	CAME	And, um, Rizzo left office and a new administration came in 'n in about eight months after th
6	HU-	HH AH1,HH AH0,HH Y UW1	HUGE	They had this, hu- +huge huge plot of land here
7	W-	W		that's how w- -- you know, how we came to be here and this organization was in existence a
8	RA-	R AE1	RATHER	Ra- +rather -- well, they started planning it prior to that. But around nineteen seventy-one c
9	A-	AH0,AE1,AA1	AS	I guess it's -- uh, uh -- I guess it's seen a- +as as the other area is west of Fifth Street. And
10	RIT-	R IH1 T		So we thought we would rit- -- we would build this first
11	FR-	F R	FROM	Probably fr- +from -- uh, through a relative, you know, his mother or father probably has so
12	F-	F,F AH0	FIRST	which sounded within our means. We -- we s- got together some thoughts on what to do to r
13	TH-	DH,DH AH0,TH	THE	And a course th- +the the river from Front to I think, uh, f- Fifth Street is the, uh, is the east
14	S-	S		we came in here after the ceremony and he told Crawford to get together with us and s- talk
15	AFFILIA-	AH0 F IH2 L IY0 EY1	AFFILIATION	affilia- +affiliation. But I'm -- I'm very
16	G-	G		And the provision was that we would raise the money -- at that time we had g- -- we had ha
17	BUILD-	B IH1 L D		Uh, build- +building -- build- -- this, this building stood but we had to rehabilitate it. It was a
18	SH-	SH	SHE	My wife said, "((Don)) have you eaten?" and he said, "No I haven't but that's alright," he sai
19	THA-	DH AE1	THAT	But in effect, he really -- I mean tha- +that the city was really doing themselves a favor. The
20				

11. Start the **forced alignment proper**~~, using the *-noprompt* option~~[4]. Don't forget to include the input file as well!

```
python FAAValign.py -vn -i inputBM.txt PH82-1-12-BMalloy.wav
```
[5]

You should now begin to see output telling you that the forced alignment is in progress (see detailed examples in the appendix).

---

[3] Please note that I had to change this setup when adapting *FAAValign* for the FAVE web site (because we did not want to allow random people to add who-knows-what to the dictionary). It is therefore **no longer the case** that additions to the dictionary will be added permanently. **You need to include your input file on every run if you want its contents to be available to the aligner!**

[4] With the "no prompt" option, you will not be prompted to confirm the transcriptions for truncated words, nor for any unknown words that are not in the dictionary (of which there should no longer be any after the procedure above).

Please note that you actually do not need to specify the "no prompt" option any longer if you are including an input file on the same run. With an input file, the program assumes that you have already checked the transcription for unknown and truncated words, and will not bother you interactively about them.

[5] The example above works if the name of the transcript file is identical to that of the sound file, e.g. `PH82-1-12-BMalloy.txt` and `PH82-1-12-BMalloy.wav`. If this is not the case, then you need to specify the name of the transcript file explicitly as the second argument:

```
python FAAValign.py -v -c unknownBM.txt PH82-1-12-BMalloy.wav PH82-1-12-BMalloy_NEW.txt
```

Appendix: Example output from the shell[6]

1. Running *FAAValign.py* with the ***–check*** option:

```
python FAAValign.py -v -c unknownMD.txt PH94-2-7-MDiPace.txt

Read dictionary from file model/dict.
Encoding is UTF-8!
Read transcription file PH94-2-7-MDiPace.txt.
Checking format of input transcription file...
       WARNING!  Empty annotation unit:  IV Interviewer  2202.105      2202.948
Checking dictionary entries for all words in the input transcription...
       Unknown word WH : Wh -- what had happened when I got married, my wife was from a
       block away. I was from sixth street, she was from seventh street..
       Unknown word ALBAN'S : Saint Alban's Street..
       Unknown word AG- : I guess the average ag- +age -- they were -- there were older
       gentlemen on the job.
       Unknown word STAMPERS : that's when a child gets mad or an adult gets mad and
       *stampers or -- or -- or.
       Unknown word FARMICOLAS : You -- you're not related to the Farmicolas or.
       Unknown word TYPIC : Uh I w- -- eh do -- would you consider yourself a typic -- a
       typical couple?.
       Unknown word CONGRATS : Congrats..
       Unknown word CHA- : That was her whole purpose. She'd been going to school and she
       wanted to make a cha- +change -- and I couldn't understand that..
       Unknown word APPREN- : You had a mentor, you had an appren- +apprentice -- an
       apprentice or some type, that you were..
Written list of unknown words in transcription to file unknownMD.txt.
```

FAAValign warns you about **empty annotation units**. This happens sometimes, usually you don't need to do anything about it.

2. Running *FAAValign.py* with the ***–import*** option:

```
python FAAValign.py -v -i inputMD.txt -c still_unknownMD.txt PH94-2-7-
       MDiPace.txt

Read dictionary from file model/dict.
Added all entries in file inputMD.txt to CMU dictionary.
Read dictionary from file added_dict_entries.txt.
Added new entries from file inputMD.txt to file added_dict_entries.txt.
Encoding is UTF-8!
Read transcription file PH94-2-7-MDiPace.txt.
Checking format of input transcription file...
       WARNING!  Empty annotation unit:  IV Interviewer  2202.105      2202.948
Checking dictionary entries for all words in the input transcription...
Written list of unknown words in transcription to file still_unknownMD.txt.
```

All entries in the input file are added to a (temporary, internal) copy of the CMU dictionary, and appended to the file `added_dict_entries.txt`.

No output between these two lines means that there are no longer any unknown words in the transcript.

[6] Lines in **bold face** represent **commands typed by the user**; everything else is shell output.

3. Running *FAAValign* for alignment:

The **–v** option produces verbose output.
The **–n** option does not prompt the user for unknown words.

```
python FAAValign.py -vn -i inputR.txt PH10-1-2-Raymond.wav

Read dictionary from file model/dict.
Added all entries in file inputR.txt to CMU dictionary.
Read dictionary from file added_dict_entries.txt.
Added new entries from file inputR.txt to file added_dict_entries.txt.
Encoding is UTF-8!
Read transcription file PH10-1-2-Raymond.txt.
Checking format of input transcription file...
Checking dictionary entries for all words in the input transcription...
Checked temporary directory .
Generated main TextGrid.
Duration of sound file:  1827.132000 seconds.
Processing Raymond -- chunk 1 :   AND THEY'RE ASKING INFORMATION ABOUT ((xxxx)) LIKE THE
       NEIGHBORHOOD LIKE AND THEY W- WANTED TO INTERVIEW ME
       Sound chunk PH10-1-2-Raymond_Raymond_chunk_1.wav successfully extracted.
       Forced alignment called successfully for file PH10-1-2-
       Raymond_Raymond_chunk_1.wav.
       Offset changed by 0.0 seconds.
       Successfully added PH10-1-2-Raymond_Raymond_chunk_1.TextGrid to main TextGrid.
Processing Raymond -- chunk 2 :   IS IT ON
       Sound chunk PH10-1-2-Raymond_Raymond_chunk_2.wav successfully extracted.
       Forced alignment called successfully for file PH10-1-2-
       Raymond_Raymond_chunk_2.wav.
       Offset changed by 28.35 seconds.
       Successfully added PH10-1-2-Raymond_Raymond_chunk_2.TextGrid to main TextGrid.
Processing Raymond -- chunk 3 :   MM YES
       Sound chunk PH10-1-2-Raymond_Raymond_chunk_3.wav successfully extracted.
       Forced alignment called successfully for file PH10-1-2-
       Raymond_Raymond_chunk_3.wav.
       Offset changed by 39.48 seconds.
       Successfully added PH10-1-2-Raymond_Raymond_chunk_3.TextGrid to main TextGrid.

 [...]

Processing Dad -- chunk 638 :   {LG}
       Sound chunk PH10-1-2-Raymond_Dad_chunk_638.wav successfully extracted.
       Forced alignment called successfully for file PH10-1-2-Raymond_Dad_chunk_638.wav.
       Offset changed by 469.663 seconds.
       Successfully added PH10-1-2-Raymond_Dad_chunk_638.TextGrid to main TextGrid.
Processing style tier.
Finished tidying up <IntervalTier "Raymond - phone" with 17422 intervals>.
Finished tidying up <IntervalTier "Raymond - word" with 6790 intervals>.
Finished tidying up <IntervalTier "Interviewer 1 - phone" with 2916 intervals>.
Finished tidying up <IntervalTier "Interviewer 1 - word" with 1250 intervals>.
Finished tidying up <IntervalTier "Interviewer 2 - phone" with 84 intervals>.
Finished tidying up <IntervalTier "Interviewer 2 - word" with 41 intervals>.
Finished tidying up <IntervalTier "Mom - phone" with 1282 intervals>.
Finished tidying up <IntervalTier "Mom - word" with 546 intervals>.
Finished tidying up <IntervalTier "Dad - phone" with 85 intervals>.
Finished tidying up <IntervalTier "Dad - word" with 36 intervals>.
WARNING!!!  Overlapping intervals <Interval "G" 0.000000:5.230000> and <Interval "sp"
       0.000000:0.000000> on tier style!!!
Finished tidying up <IntervalTier "style" with 698 intervals>.
WARNING!  Overlapping intervals detected!
Error messages saved to file PH10-1-2-Raymond.errorlog.
Successfully written TextGrid PH10-1-2-Raymond.TextGrid to file.
Written log file PH10-1-2-Raymond.FAAVlog.
```

This is where the alignment proper starts, breath group by breath group.

If the forced alignment produces a TextGrid with **overlapping intervals**, *FAAValign* prints a **warning message** and writes the information about the overlapping intervals to a `.errorlog` file.