Annotation guidelines
Chamorro Psycholinguistics Project
Elicitation Experiment

Jake Vincent

5/6/2016

1 Basics

1.1 The experiment

The data that we'll be working with came from an elicitation experiment conducted by Matt Wagers, Sandy Chung, and Manny Borja in the Mariana Islands. The goal of the experiment was broadly to elicit relative clauses. Participants were shown cards with two pictures each (see `scenes.pdf` file). The cards were labelled with an uppercase letter to indicate the trial number. The two pictures on a single card differed in two ways from each other. For the card labelled A, for example, the experimenter would ask "Which woman is tickling the boy?" and the participant would hopefully respond with a sentence containing a relative clause or two, e.g. "The woman who is tickling the boy is the one who is wearing the woven hat."

1.2 Getting started

The program we'll be using to annotate the audio files from the elicitation experiment is called **Praat** (http://www.fon.hum.uva.nl/praat/). Praat contains a useful tool for annotating audio files directly called "TextGrid." TextGrids are highly structured text files that Praat saves separately from the audio file but that have the same name as the audio file. When the TextGrid and audio file are opened together in Praat (select the audio and TextGrid file at the same time in Praat's "Objects" pane, then click `View & Edit`), a large window opens with the waveform and spectrogram in the top pane and the annotation information in the lower pane.

In Praat, it is possible to annotate on multiple "tiers." This allows you to have multiple levels of information associated with the audio file, and allows those different levels to overlap each other in the time course of the audio file. Praat is capable of creating two different kinds of tiers—**point tiers** and **interval tiers**. Interval tiers allow you to annotate a certain *interval* of time, while point tiers only allow you to annotate a point in time, which might matter if you were only concerned about the onset of certain things. **We will only be using interval tiers**.

For our purposes, all of the audio files from the experiment already have an associated TextGrid file that you'll see in the shared folder. If no one has worked on the annotation for that file since they were initially created, then the TextGrid will only have four tiers: `trial`, `answer`, `question`, and `comment`. The `trial` and `answer` tiers should already be filled in with the trial letter associated with the picture. On the `trial` tier, the whole question-answer period will be the selected interval, and on the `answer` tier, only the participant's response will be the selected interval.

2 The tiers

In this section, first we'll go over the names of the tiers we want to use, how to create and add information to these tiers, and what kind of information should go in them. **We will be enriching the already-created TextGrids by adding more interval tiers**. What we want to end up with is a TextGrid with **ten** tiers labelled exactly as follows and in this exact order (you'll notice that the four bottom tiers are the ones that were already present in the TextGrid file you opened):

1. `word.text`
2. `word.gloss`
3. `animacy`
4. `RC`
5. `disfluency`
6. `translation`
7. `trial`
8. `answer`
9. `question`
10. `comment`

To add these new interval tiers to the TextGrid file, make sure you already have the TextGrid interface open. Click `Tier` in the menu at the top, and in the dropdown menu, click `Add interval tier`. A window will open that will allow you to give the desired position for the tier, and the name for the tier. I'd recommend starting at the top of the list given above. Follow this procedure:

- Add the `word.text` tier
  - Open the `Add interval tier` interface as instructed above
  - In the `Position` field, clear what was there and enter `1`
  - In the `Name` field, enter `word.text`
  - Click `Ok`

- Add the `word.gloss` tier
  - Open the `Add interval tier` interface again
  - In the `Position` field, enter 2
  - In the `Name` field, enter `word.gloss`
  - Click `Ok`
- ...

Follow this procedure until you've reached the last new tier you need to add. All the previously existing tiers in the TextGrid file will get bumped down to the positions we want them to be in.

2.1 Information that goes in the tiers

When adding information to the tiers, we'll follow the simple rule of using dashes (-) to indicate segmentable or parsable material, and dots (.) to indicate unparseable or unsegmentable material. We will also adopt two additional conventions from the Leipzig Glossing Rules for some specific cases: preverbal agreement clitics will be separated from the verb root using an equals sign (=), both in the Chamorro text and in the gloss. In some specific cases discussed below, we will also surround certain elements in parentheses (( )) to indicate that it is an inherent part of another element.

`word.text` This is where the Chamorro text goes. To add the Chamorro, first reference the transcriptions that are associated with each audio file. These audio files are for the most part a word-level analysis and do not indicate morpheme boundaries within words. Our annotation will go deeper than this, so **we need to do some morphological parsing of the Chamorro text provided in the transcription file**. We need to look out for **prefixes**, **infixes**, **suffixes**, and **clitics**. Here are some guidelines:

1. **Affixes**: **separate affixes from the stem using a single dash**, e.g. *ma-tufuk* 'woven' (PASS-weave).

   Affixes to look out for:

   (a) Prefixes: passive (*ma-*), antipassive (*man-*), plural (*man-*), causative (*na-*), and many others

   (b) Infixes: *-in-* (many functions), *-um-* (various kinds of agreement)

   (c) Suffixes: possessor suffixes (*-hu, -mu, -ña*, etc.), the post-head form of the linker (*-n*), and others

2. **Clitics**: **separate clitics from the stem using an equals sign** (=), e.g. *hu=taitai i lepblu* 'I read the book' (1SG.AGR=read DEF book). **For our purposes, we are only going to consider the preverbal person-number-mood agreement elements clitics**.

`word.gloss` This is where the English glosses will go. On this tier we are going to adhere to some more specific rules and conventions.

First of all, **make sure there is a one-to-one ratio of morphemes for the Chamorro text and their corresponding glosses**. If you parse the word meaning 'woven' into two morphemes (*ma-tufuk*), make sure that each morpheme has its own gloss; instead of putting 'woven' into the `word.gloss` tier, put 'PASS-weave'.

When you find an infix, place the English gloss for that infix before the gloss corresponding to the root. For example, if you encounter a word with the passive infix, place the PASS gloss before the gloss for the root. For example, the root *kassi* means 'tease', but can take the passive infix *-in-*, as follows: *k-in-assi*. For cases like this, you'll gloss them 'PASS-tease'. If the word has agreement clitics or other prefixes, still make sure that the gloss for the infix is closest to the root.

For the purposes of this study, we want to **indicate the syntactic categories of the words in the `word.gloss` tier**. The category will be the first component of the annotation in this tier, and we will place them in parentheses (following the Leipzig glossing rules) to indicate that they are an inherent part of the word. Follow these conventions for marking syntactic category:

1. Nouns = (N)
   (a) common nouns, e.g. *kareta-ña* = (N)car-3S.POSSR
   (b) names, e.g. *Maria* = (N)maria
2. Determiners = (D)
   (a) definite determiner (*i* 'the')
   (b) indefinite determiner (*un* 'a')
   (c) demonstratives (*atyu, etyu* 'that', *esti* 'this')
3. Verbs = (V)
4. Adjectives = (A)
5. Adverbs = (ADV)
6. Quantifiers = (Q)

`animacy`

`RC`

`disfluency` Disfluencies include anything in a segment of speech that is not fluid, normal-paced speech. Here we'll be referencing the disfluency types indicated in a manual for the CHILDES project, a project for studying child speech. We will probably encounter several different kinds of disfluency in this annotation project. Here is a list of disfluency types that we may encounter, and what they mean:

1. **Fillers** will be annotated as `DISFL(filler)`.
2. **Pauses** will be annotated as `DISFL(pause)`.
3. **Drawling** or **lengthening** will be annotated as `DISFL(length)`.
4. **Repetitions** will be annotated as `DISFL(repeat)`.
5. **Retracings** will be annotated as `DISFL(retrace)`.

fillers initial repetition blocking internal pausing drawling, lengthening word repetitions retracings actual words repeated actual words retrace

`translation`

`trial`

`answer`

`question`

`comment`