Anaphora Coding Protocol

Brief overview of anaphora

Anaphora are expressions that refer to other objects or entities introduced earlier in a discourse to avoid repetition (Mitkov, 2014). The interpretation of an anaphor is determined by the interpretation of the **antecedent**, the entity that the anaphor refers to (Mitkov, 2014; Lust, 1981). There are several types of anaphora that are commonly used in natural language, below are some of the most common:

- **Pronominal anaphora:** anaphora that use pronouns. Example:

  > "Computational Linguists from many different countries attended the tutorial.
  >
  > They took extensive notes." (Mitkov, 2014)

  Note that not all pronouns are anaphoric (e.g. "It is important"). A non-anaphoric "it" is called **pleonastic**.

- **One anaphora:** using the word "one" to refer to the antecedent (Sukthanker et al., 2018). Example:

  > "If you cannot attend a tutorial in the morning, you can go for an afternoon one." (Mitkov, 2014)

- **Split anaphora:** a pronoun can refer to more than one antecedent (Sukthanker et al, 2018). Example:

  > Katherine and Maggie love reading. They are also the members of the reader's club." (Sukthanker et al, 2018)

Coding: Step by Step

1. After you open the speech transcription file, highlight **Column C, D, E, F** and right click your mouse, scroll down to "Insert" and insert **FOUR** columns between **Column B** and **Column C**.
2. Change the file name to `speech_#####–coded.txt`, where ##### is the participant ID.
3. Check each sentence or phase in **Column G** to see if there is an anaphor.
   a. If there is not an anaphor, leave the inserted columns blank.
   b. If there is an anaphor:
      i. In **Column C**, write the anaphor word or phrase. If there are multiple anaphora in the sentence/phrase in Column F, list all anaphora in Column C, separated by commas.
      ii. Determine what object the anaphor is referring to. In **Column D**, write the referent ID (see below).
         1. If there are multiple anaphora listed in Column C, list the reference IDs for each anaphor in Column D, separated by commas.
         2. If one anaphor references multiple objects, write the IDs separated by forward slashes. For example, if "they" refers to both the ladybug and praying mantis, write: 11/12.

      iii. In **Column E**, mark the type of anaphora:
         1. Pronominal anaphora (write: pronoun)
         2. One anaphora (write: one)
         3. Split anaphora (write: split)

      If there are multiple types of anaphora listed in Column C, list each type of anaphora in Column E, separated by commas.

iv. In **Column F**, write the **cue variable**. If the anaphora can be determined by speech only, write **1**. If the anaphora needs both speech and visuals to be determined, write **2**. If there are multiple anaphora listed in Column C, list the disambiguation variable for each anaphora in Column F, separated by commas.

1	speech only
2	speech **and** visuals required

Referent IDs

Numbers will be used to identify each of the toys in study. Images of the toys can be found in the anaphora project Google Drive folder in `anaphora project/toys`. For the most part, the only objects/entities that will be referenced in the dialogues by anaphora will be from the list below:

Toy name	**ID**
helmet	1
house	2
blue car	3
flower	4
elephant	5
snowman	6
rabbit	7
SpongeBob block	8
turtle	9
hammer	10
ladybug	11
praying mantis	12
green car	13
saw	14
doll	15
phone	16
Rubik's Cube	17
rake	18
truck	19
white (police) car	20
spinning drum (ladybug)	21
purple block	22
bed	23
beach ball block	24
people	25
non-study objects	26

Each toy corresponds to one unique ID value, which will be used when coding every transcript file.

Below are a few notes on special cases:

- **People:** if the child or the parent is referenced, use the ID number **25**. If other people not in the room are referenced, do *not* code for then.
- **Non-present objects:** do *not* code for referents that are not physically present in the room.
- **Additional objects:** for objects that are referenced that are not explicitly part of the study but still present in the room, use the ID number **26**.

Summary

- **Column C:** anaphora expression(s)
- **Column D:** referent ID(s)
- **Column E:** anaphora type(s)
- **Column F:** cue variable(s)