Metrics for assessing the representations underlying children's productions of multi-adjective strings

Lisa S. Pearl

1 Representational hypotheses

These are possible hypotheses for the representations that underlie adjective ordering preferences in child-produced speech, in particular multi-adjective strings of the form $adj_2\ adj_1\ noun$ like "*cute grey kitty*".

1. **Perceived subjectivity** (H_{Subj}): By individual adjective (i.e., subjectivity score = a real-valued number between 0 and 1) or by class (binned by SDG perceived subjectivity over lexical semantic classes: LOW, MIDDLE, and HIGH)
2. **Lexical semantic class** (H_{SemCl}): Taken from SDG's synthesis of the literature supporting this (e.g., Dixon 1982, Cinque 2014) DIMENSION, VALUE, AGE, PHYSICAL, SHAPE, COLOR, and MATERIAL.
3. **A baseline** ($H_{InputFreq}$): Frequency in input (i.e., child-directed speech) of adjective in second position.

2 Metrics

2.1 Comparing against average observed distance from noun

This is what we currently calculate in our plots: the average distance of the adjective from the noun (0 = next to the noun (adj_1), 1 = one position away from the noun (adj_2)).

The frequency that a particular adjective adj_x occurred in the second position of $N_{prod}(adj_x)$ child-produced multi-adjective strings containing adj_x is $f_{2prod}(adj_x)$. This means the observed probability of adj_x appearing in the second position of a multi-adjective string is

$$p_2obs(adj_x) = \frac{f_{2prod}(adj_x)}{N_{prod}(adj_x)} \quad (1)$$

The expected probability of adj_x appearing in the second position ($p_2exp(adj_x)$) depends on the representational hypothesis. This should match the observed probability if the representational hypothesis is a good match for what's really going on (i.e.,

$p_2obs(adj_x) = p_2exp(adj_x)$). Let's consider each of our three hypotheses in section 1 in turn.

In common: We're interested in the probability that adj_x appears in the second position when in combination with some other adjective (adj_{oth}) the child has available.

Intuitions:

1. H_{Subj}: adj_x will appear in the second position if adj_{oth} is less subjective than adj_x. If the subjectivity is the same for both adjectives, then adj_x has a 50% chance of appearing second.

2. H_{SemCl}: adj_x will appear in the second position if adj_{oth} is in a lexical semantic class that appears closer to the noun than adj_x's lexical semantic class. If the lexical class is the same for both adjectives, then adj_x has a 50% chance of appearing second.

3. $H_{InputFreq}$: adj_x will appear in the second position as often as it appeared in the second position in the child's input (irrespective of what adj_{oth} was).

How can we operationalize these intuitions? One way:

1. H_{Subj}: Determine the number of adjective tokens in the child's input appearing in multi-adjective strings that have a lower subjectivity score than adj_x ($f_{input}(< adj_x)$). Determine the number of adjective tokens in the child's input that have the same subjectivity score as adj_x ($f_{input}(= adj_x)$). Determine the total number of adjective tokens appearing in multi-adjective strings in the child's input ($N_{input}(adj)$).

$$p_2exp(adj_x) = \frac{f_{input}(< adj_x) + 0.5 * f_{input}(= adj_x)}{N_{input}(adj)} \quad (2)$$

   For example, in the child's input, suppose there are 200 adjective tokens appearing in 100 multi-adjective strings ($N_{input}(adj)$=200), of which 50 have a lower subjectivity score than adj_x ($f_{input}(< adj_x)$=50), 5 have the same subjectivity score ($f_{input}(= adj_x)$=5), and 145 have a higher subjectivity score. $p_2exp(adj_x)$=$\frac{50+0.5*5}{200}$ = 0.2625.

2. H_{SemCl}: Determine the number of adjective tokens in the child's input appearing in multi-adjective strings that have a closer semantic class than adj_x ($f_{input}(< adj_x)$). Determine the number of adjective tokens in the child's input that have the same lexical class as adj_x ($f_{input}(= adj_x)$). Determine the total number of adjective tokens appearing in multi-adjective strings in the child's input ($N_{input}(adj)$). $p_2exp(adj_x)$ is calculated the same as equation (2) above.

   For example, in the child's input, suppose there are 200 adjective tokens appearing in 100 multi-adjective strings ($N_{input}(adj)$=200), of which 50 are in a closer lexical semantic class than adj_x ($f_{input}(< adj_x)$=50), 5 are in the same lexical semantic class ($f_{input}(= adj_x)$=5), and 145 are in a semantic class that's farther from the noun. $p_2exp(adj_x)$=$\frac{50+0.5*5}{200}$ = 0.2625.

3. $H_{InputFreq}$: Determine the number of multi-adjective strings in the child's input that contain adj_x ($N_{input}(adj_x)$). Determine the number of those where adj_x occurs in the second position ($f_{2input}(adj_x)$).

$$p_2exp(adj_x) = \frac{f_{2input}(adj_x)}{N_{input}(adj_x)} \tag{3}$$

For example, in the child's input, suppose there are 50 multi-adjective strings containing adj_x. Suppose that in 20 of these, adj_x appears in the second position. $p_2exp(adj_x) = \frac{20}{50}$ = 0.40.

For each hypothesis, we can calculate $p_2obs(adj_x)$ and $p_2exp(adj_x)$ for each adjective in the child-produced dataset, and then run a correlation to see which hypothesis is the best fit.

2.2 Comparing different representational hypotheses

This is basically model comparison, where we calculate the likelihood (=probability of the observed child-produced speech data, given the generative hypothesis). This can then be used to compute a Bayes factor, which is just the ratio of one generative hypothesis's likelihood to another.

So, no matter what the representational hypothesis, we want to calculate the probability of the observed multi-adjective data produced for adj_x, given that hypothesis (i.e., $p(D(adj_x)|H)$). This is the likelihood.

To do this for a given hypothesis H, we need to use the quantities we previously used to calculate $p_2obs(adj_x)$ in equation (1): the frequency of adj_x in the second position when it appeared ($f_{2prod}(adj_x)$) and the total number of multi-adjective strings produced that contained adj_x ($N_{prod}(adj_x)$). Let's abbreviate these quantites with f and N, respectively. The probability of producing the observed f instances of adj_x in second position, given N multi-adjective string samples with adj_x is

$$p(D(adj_x)|H) = \binom{N}{f}(p_2exp(adj_x))^f(1 - p_2exp(adj_x))^{N-f} \tag{4}$$

We then do this for each adjective, where the likelihood for all adjectives $adj_x \in A$ in the child-produced dataset ($p(D|H)$) is

$$p(D|H) = \prod_{adj_x \in A} p(D(adj_x)|H) \tag{5}$$

We can then compare $p(D|H)$ for each representational hypothesis against one another. For example, to compare H_{Subj} to H_{SemCl}, we would calculate $\frac{p(D|H_{Subj})}{p(D|H_{SemCl})}$ and get an official Bayes factor.

Note: Because the probabilities may get super-small when we're multiplying things together, we probably want to calculate $p(D|H)$ in log space (see equation (6) below), and then un-log the quantities for the Bayes factor ratio.

$$log(p(D|H)) = \sum_{adj_x \in A} log(p(D(adj_x)|H)) \tag{6}$$

$$Bayes\ factor = \frac{e^{log(p(D|H_1))}}{e^{log(p(D|H_2))}} \tag{7}$$

Another way to do this is to keep in mind that we actually want $\frac{p(D|H_1)}{p(D(H_2)}$:

$$Bayes\ factor = \frac{p(D|H_1)}{p(D|H_2)} = e^{log\ \frac{p(D|H_1)}{p(D|H_2)}} = e^{log(p(D|H_1)) - log(p(D|H_2))} \tag{8}$$