Closer look into formulas

1. **Attention distribution (over source positions):**

$$e_i^j = w^T \tanh(W_h h_i + W_s s_j + b_{attn})$$

$$p^j = softmax(e^j)$$

2. **Vocabulary distribution (generative model):**

$$v_j = \sum_i p_i^j h_i$$

$$p_{vocab} = softmax(V'(V[s_j, v_j] + b) + b')$$