Data:

The data set for these rules were all the sentences with “be” from the SPOHP oral interviews. Each sentence was annotated by human annotators in order to train the model. Later, additional, data augmented sentences were added to the data set in order to balance it. These sentences were based on habitual sentences in the original data set, and then run through a filter such that they were most likely habitual and thus labeled as such.

Rules:

A rule tagging a sentence means it fits the requirements of that rule, making it true. If it is true, it is marked as 1, if it is false, it is marked as 0.

The rules were parsed using SpaCy’s dependency parser.

1) *POS1*

The word immediately preceding “be” is a modal, an adjective or “to”, most commonly “to” preceding “be”, then modal preceding “be” and finally adjective preceding be

e.g. “Well, no, what you mean what as that like to born to be born in a…”
e.g. “But he didn't prepare properly so that it could be continued.”

1) POS2

The word immediately following “be” is an adjective and the word preceding be is not a personal pronoun or noun

e.g. “He told him whatever he did, if he didn't get to go to college, be good at whatever he did.”

2) POS3

The word immediately following “be” is a preposition or subordinating conjunction and the word immediately preceding “be” is a singular present verb

e.g. “When you say be mindful of what do you mean?”

3) POS4

The word preceding “be” is a noun and the word preceding that noun is an adjective.

e.g. “And I ask the lord that my last day be my best.”

4) POS5

The word preceding “be” is an adverb and the word following “be” is either a personal pronoun or determiner.

e.g. “I'm only here to teach, not be a baby-sitter.”

5) POS6

The word preceding “be” is an adverb and the word preceding the adverb is a verb or modal.

e.g. “It can't be no worse than it has been!”

6) A1

The 'be' is preceded by don't and the word before 'don't' is a noun or pronoun.

eg. She don't be working out.

Or

The 'be' is preceded by a word that is not a verb or auxiliary, which is preceded by a 'don't' and a noun or pronoun.

eg. They don't really be talking

7) A2

Commented [WP1]: Ad-Hoc non-Habitual

The 'be' is followed by certain parts of speech; tends to be Non-Habitual. These parts of speech include:
INTJ : Interjection
CCONJ: Conjunction
DET: Determinant
PROPN : Proper noun
PUNCT: Punctuation

eg. It wasn't as big as I thought it would be.
eg. It going to be an interesting time to wonder. ('an' is a Determinant)

8) A3

Commented [WP2]: Ad-Hoc Habitual; finds be's that are directly preceded by a pronoun

(the other rules already capture this)

The 'be' is directly preceded by a pronoun or indirectly preceded by a pronoun and the words between the pronoun and 'be' are not auxiliaries , verbs or particles; tends to be Habitual.

eg.I just be liking the beat to a hip hop song.
Here, 'be' is indirectly preceded by a pronoun and 'just' here is an adverb.(which is not AUX or Verb or Particle)

eg.I be listening to the beats

9) A4

Commented [WP3]: Ad-Hoc rule; habituality

'Be' is followed by a verb ending in 'ing' and is not preceded by an auxiliary verb, ‘to’, or any of the words in phonetic variation: ‘gonna’, ‘gotta’, ‘wanna’, or ‘tryna’; tends to be Habitual.

eg. Elysa be showing me some work

10) A5

Commented [WP4]: Ad-Hoc Non-habituality

'Be' is preceded by a word ending in ‘n't’ which is not ‘don't’; it tends to be Non-Habitual.

eg. I mean, you can but you wouldn't be too successful with it

11) SynPar1

The “be” in the sentence has children with an aux dependency relation and an AUX upos tag

e.g. “All the things that children should do for parents we did, I did, and she says it will be a different day and further up on that road, you're gonna see what I am talking about.”

12) SynPar2

The “be” in the sentence has siblings with an aux dependency relation and an AUX upos tag

e.g. “Those barriers may not really be holding them back and they could possibly find a way to do something about the issues that exist.”

13) SynPar3

The “be” in the sentence has an aux dependency relation and an AUX upos tag and it’s head has a upos tag of VERB

e.g. “Like a person be lying.”

14) SynPar4

The “be” in the sentence is labeled has a VERB upos tag

e.g. “We had study hour, and from that, there was no reason for you be out and stuff.”

15) R1

If none of the pos rules are true, the ‘be’ tends to be Habitual.

Rule Interactions:

Rule interactions help make the rules stronger; i.e. lean more toward habitual or non-habitual. For example, a rule (rule 1) might somewhat indicate habituality, unless another rule (rule 2) is true, in which case the be is non-habitual. By setting rule 1 to 0 if rule 2 is true, rule 1 will be a better predictor of habituality.

e.g. if (Sentence.a4 == 1):
Sentence.a5 == 0

In the above, a4 indicates a habitual be, while a5 indicates a non-habitual be, however if both are true, the sentence is usually habitual. Thus, a5 is marked 0 to reflect this idea. In fact, observing the training data, every sentence still marked with a5 after applying this interaction is non-habitual.

Full list of the rule interactions

```
for Sentence in sen:
    if (Sentence.r1 == 0):
       Sentence.synPar3 = 0
    if (Sentence.a4 == 1):
       Sentence.synPar3 = 0
    if (Sentence.a2 == 1 and Sentence.a3 == 1):
       Sentence.pos5 = 0
    if (Sentence.a4 == 1):
       Sentence.a5 == 0
    if (Sentence.a3 == 1):
       Sentence.synPar2 = 0
    if (Sentence.pos6 == 1 or Sentence.synPar1 == 1):
       Sentence.a3 = 0
    if (Sentence.synPar1 == 1):
       Sentence.r1 = 0
```

N-grams:

First, the tagger is shown a dataset where it observes the four words before and after each ‘be’ and the habituality tag of each ‘be’. A MultinomialNB model is trained on this data and exported. Using the information it has learned, the model then predicts the habituality of new sentences from the four words before and after the ‘be’ in each sentence. Finally, this initial prediction is used as one of the variables to ultimately determine habituality.

POS:

Certain POS (part-of-speech) tags are taken into account by many of the habituality rules. However, in addition to this, there are numerous patterns that may not have been considered by the rules. Thus, the POS tags are also used as separate variables to help the tagger determine whether each 'be' is Habitual or Non-Habitual.

Model:

The habituality model was created with all the inputs above as the predictor (X) values used to predict the predicted (y) value. To train the model and assess its accuracy, we using a stratified k-fold ensemble model. The stratified k-fold component divides the data into 10 equal, balanced datasets and rotates which 9 are the train and 1 is the test. The ensemble model combines the results of the logistic regression (lr), multilayer perceptron (mlp), and support vector classification (svp) models.
The ensemble model is run on each of the train/test split combinations and the results are averaged to a percentage.

The higher the percentage for each sentence, the more confident the model that the 'be' is non-habitual. Based on this, the 'be' is classified as -1 or 1, -1 being non-habitual and 1 being habitual. The default threshold value is 0.5, that is, above the threshold means non-habitual, and below means habitual. However, we adjusted the threshold to 0.84 to increase the recall of the Habitual class. The reasoning here is that with very high habitual recall, annotators only have to verify that the sentences tagged habitual are in fact habitual; the sentences tagged non-habitual will almost always be tagged correctly and thus can be accepted as is. In addition, there are many fewer habitual 'be' in most transcripts as compared to non-habitual 'be', so this even further reduces the amount of manual verification that needs to be done after automatic tagging

Regarding the data, at first, there were many fewer habitual sentences than non-habitual, making it more difficult for the model to accurately tag sentences as habitual. Thus, we ran the habitual sentences through data augmentation code written by Harrison Santiago. This resulted in us having a roughly equal number of habitual and non-habitual sentences for the model to be trained on.

In the case of multiple be's, the model is run on each individual 'be', then the results of each 'be' are combined together, e.g. "1, -1, 1", indicating the first and last be are habitual but the middle be is not.

Key/Definitions

- 1: habitual be, -1: non-habitual be, 0: no be
    - If multiple be's in a sentence, tags each be, e.g. "-1, 1"
- Stratified k-fold – rotates train and test data while preserving the percentage of samples in each class (e.g. 40% non-habitual, 40% habitual), then averages results
- Ensemble – mix of results from lr, mlp, svc
- Pipeline – standardizes data
- Threshold – higher threshold = more likely to class 'be' as habitual, thus increasing the recall for the Habitual class

Results

Below is the classification report of the Habituality model. Initially, the threshold for classing 'be' as non-habitual was 0.94, that is, the tagger had to have 94% certainty or higher that the 'be' was non-habitual in order to class it as such. As a result, the recall for the Habitual class was very high (around 0.99), which seemed ideal for annotating. However, in practice, the 0.94 threshold resulted in excessive amounts of non-habitual sentences predicted as habitual without a clear benefit in regards to accurately predicting habitual sentences as habitual. Thus, we chose a lower threshold of 0.84 for better results.

```
In [100]: confusion_matrix(y, prediction_new)
Out[100]: array([[3568,  446],
                 [  84, 3641]])

In [101]: print(classification_report(y, prediction_new, target_names=target_names))
                precision    recall  f1-score   support

     nonhabitual       0.98      0.89      0.93      4014
        habitual       0.89      0.98      0.93      3725

        accuracy                           0.93      7739
       macro avg       0.93      0.93      0.93      7739
    weighted avg       0.94      0.93      0.93      7739
```

As you can see above, the accuracy, recall, and precision are all quite high, particularly the recall for habitual sentences and precision for nonhabitual sentences.

We output the components of this model as cv.joblib, ngram.joblib, and habituality_model.joblib. Then, we wrote a second script to analyze each sentence to generate the inputs like rule truth values, pos value, etc. Finally, we ran the model on each file and adjusted for the threshold of 0.84.