All documents are processed using BookNLP. BookNLP is a natural language processing pipeline that performs tokenization, pos-tagging, ner-tagging, dependency-parsing and many other operations. Some BookNLP outputs -- such as NER tags, supersense tags, dependency tags etc. -- are utilised in the computation of the features described below. We use three larger categories, tense/mood/voice, to group our features, which are based on Genette's triangle of relations between story, discourse, and narrating (1983). A fuller description of narrative theory is provided in Piper, So, and Bamman (2021).

Tense

This category is concerned with capturing the relations between *story* and *discourse.* It consists of one feature:

1. Temporality: Temporality is designed to capture the explicit invocation of temporal frameworks within narrative discourse. We hypothesize that narrative texts or texts scoring higher in narrativity will exhibit higher rates of temporal discourse. Because actual grammatical tense is captured through POS tags, here we condition on the following NER tags related to time:
    a. Numerator → Number of "TIME", "DATE", "DURATION", "SET" *ner* tags
    b. Denominator → Number of words (excluding punctuation)

Mood

This category captures the relationship between eventfulness, i.e. "what happens," and descriptiveness, i.e. the construction of a reality. It includes four features:

2. Setting: We hypothesize that the more situational details are provided, i.e. the more the environment of events are explicitly invoked, the more we can see Herman's idea of "world-making" function of narrative being performed (Herman 2009). We measure *setting* as a combined construct of thing-places, i.e. the rate of geographic places, physical locations, and artifacts relative to the total number of words of a passage. For geographic places, we use the NER location tags. For physical locations, we crawl the hyponym trees of certain seed words from Wordnet (see /src/classifier/places-hyponyms.ipynb). And for artifacts, we use the supersense tag "noun.artifact." Therefore, *setting* is computed as a ratio of:
    a. Numerator → Number of "LOCATION" ner tags + Number of "noun.artifact" supersense + Number of "ORGANIZATION" ner tags (with pobj) + Number of matches from a predefined places-vocabulary (created via hyponym tree traversal)
    b. Denominator → Number of words (excluding punctuations)

3. Concreteness: Herman's (2009) emphasis on "lived experience" and Bruner's (1991) distinction that narrative's opposite is more conceptually driven, reason-based communication suggests that narrativity should exhibit higher levels of concrete language (porches, houses, fields, trees) than abstract language (principles, rationality, consciousness, etc). We use Brysbaert et al's lexicon of concretization as a means of capturing a passage's degree of concreteness and particularity versus its conceptuality and generality. It is computed as a ratio of:
    a. Numerator → Sum of concreteness scores of every word/bigram in the text that appears in their lexicon.
    b. Denominator → Number of words (excluding punctuations)

4. Saying: Acts of speech are narrative ways of drawing attention to the in-the-worldness of information (very close to the perspectival language of voice). We thus capture the dimension of *saying* as explicit invocations of communication, i.e. we do not differentiate between direct or indirect speech, nor do we measure the extent of dialogue itself, but merely the referencing of the act of communication itself. It is a ratio of:
    a. Numerator → Number of words with “verb.communication” as supersense
    b. Denominator → Number of words (excluding punctuations)

5. Eventfulness: We hypothesize that narratives will exhibit higher rates of events than non-narratives because of their emphasis on experientiality, which is also another form of concretization. We measure *eventfulness* as the rate of all non-helping and non-stative verbs in a passage relative to the number of words:
    a. Numerator → Number of non-helping and non-stative verbs -- using a predefined set of helping verbs (see src/classifier/config.py) and “verb.stative” supersense
    b. Denominator → Number of words (excluding punctuations)

Voice

Voice can be conceptualized as a function of the explicitness of an agent's experientiality, the "lived" aspect of "lived experience" in Herman's terms (2009). This category captures things related to agents and consists of four features:

c. Feltness: We hypothesize that narrativity is positively correlated with increased levels of perceiving, cogitating and affective agents. It is measured as the rate of perceptual, cognitive, or emotion terms over the total number of words:
    i. Numerator → Number of words with noun.cognition or verb.cognition or noun.feeling or verb.emotion or verb.perception as bookNLP supersense
    ii. Denominator → Number of words (excluding punctuations)

6. Agenthood: Since narrative is focused on lived experience by agents told from a particular agent's point of view, we hypothesize that the predominance of animate agents

will be indicative of higher levels of narrativity. We model agenthood as the presence of an "animate entity" in the subject position of a sentence. Animate entities are detected using a combination of (i) ner ‘PERSON’ tag and (ii) supersense 'person/animal' tag. If one (or more) of these methods consider a token animate, it’s labeled as animate. Therefore, *agenthood* is a ratio of:

   a. Numerator → Number of "animate” entities in the nsubj position
   b. Denominator → Number of words (excluding punctuations)

7. Agency: While agenthood captures the prevalence of animate subjects, agency captures the proximity of such subjects to eventful actions. In other words, the more tightly knit subjects are with their actions, the more agency these subjects are grammatically endowed with. We calculate the average distance between a sentence's subject and the most immediate verb (measured in number of tokens) and then compute the mean for all 'nsubj' tokens in the passage.

8. Coherence: According to Herman's theory of event sequences, we also expect to find higher levels of continuity with respect to the entities present in a text, which we might also understand as a measure of "focalization," where the sequential experience of a particular agent is foregrounded. To measure coherence, we calculate the rate at which entities recur from one sentence to the next (‘seq’) and within a larger text span (‘global’) relative to the total number of entities. We match entities through word repetition and co-reference resolution. It is computed as the average number of token-overlap for all pairs of sentences (‘global’) and sequential pairs of sentences (‘seq’) in the given passage.
   a. seq looks at sequential pairs of sentences. In a 5-sentence passage, this corresponds to indices 0-1, 1-2, 2-3, 3-4
   b. Computes the number of tokens that overlap in sent 1 and sent 2: iterates over tokens in the first sentence:
      i. +1 if it appears in the second sentence
      ii. +1 if its coreferent appears in the second sentence.
      iii. Only looks at relevant words -- nouns, verbs, adverbs, adjectives and pronouns (via coreference resolution).
      iv. Additionally, when char_id for a PERSON in the first sentence is -1 (that is, BookNLP failed at coreference resolution), we count it as an overlap with a pronoun (if exists) in the second sentence.

Lastly, we also use the feature *Pct-Quoted*: this is computed as the ratio of the number of words in quotes to the total number of words (excluding punctuations). BookNLP’s *Quotation* label is used to determine whether a given word is inside quotation marks or not.