FINAL STRATEGY

# STRATEGY:
## Calculating Cognitive Informativity (CI) of text translation:
### Mathematical model:

CI = (Word Translation(WT) + Meaning Translation(MT) + Register Translation(RT) + Aesthetic Translation(AT)) / 4

A weighted average can also be taken while finetuning the model

Suppose J is the original language and S is the language of the translated text

WT = Semantic Similarity between V(T) and V(S): this can be found out using WordNet for words of the same language but for words of a different langauge, strategy to compute is explained later

MT = lexical similarity between V(T) and V(S): explained in detail later (to be worked upon more)

RT = tone or emotion or style : eg formal, casual, intimate, positive, negative when comparing V(T) and V(S): Sentiment Analysis: can be done using SOTA sentiment analysis APIs

AT = (WT + MT + RT)/3

### To compute Semantic Similarity:

Use dependency parser on 2 sentence in different languages and then using it get word(pairs) with the same parts of speech in both the original text and th translated text.

For these pair of words use a standard dictionary to translate the words to a same language and then get semantic similarity which will a weighted average of the pairs based on the part of speech.

### Extending the single sentence idea to paragraphs:

Assumption: If the position of a sentence in one language gets shifted to more than 3 sentences away from its original positioning in the translated language then semantic similarity should go down. Also breaking a sentence into 2 or more parts in the translated text should lower the semantic similarity.

Hence for each sentence, parse the sentence at the same location as in the original text in the translated text and a sentence before it and a sentence after it to find the most matching sentence using dependency parser/ embeddings and now find the sentence similarity score as explained above.

In the case of a broken sentence (generally rare), the most matching sentence would be found but as it won't contain the entire information of the sentence a lower informative similarity score would be assigned which is justified as the sentence is broken into 2 or more parts while translation which affects the readibility of the text.

### Future goals:
- explore more about knowledge graphs
- explore more about embeddings