N grams are sequences of words that appear in documents. N-grams are important in NLP because of their extensive applications. These sequences of continuous words the n refers to the number of words in the sequence. There are unigrams for 1 word sequences, bigrams for 2 word sequences, trigrams for 3 word sequences, and anything larger would generically be called an n-gram. They can be used to make meaning of large texts, if a sequence appears a lot of times in a large text then you can use those instances to understand or find greater meaning. For calculating the probabilities for bigrams you use (b+1)/(u+v) where b is the bigram count, u is the unigram count of the first word in the bigram, and v is the total vocabulary size (add the lengths of the 3 unigram dictionaries). The reason source text is so important in building a language model is due to the fact that we need examples of that language in use to build a model. With source texts we have data to train and test our model on to further refine and develop our language models. Without that data we wouldn't be able to create an effective model. Smoothing is important so that you have a non-zero probability for words that haven't been seen yet so that they can still be predicted by the model in the rare case where such word does appear. It proves the estimation for a given word if it never appears in the set. Using language models for text generations has been done this is because you can predict the probability of certain words appearing in a sequence using n grams. The limitations of such technology are that the text generation might not be able to understand context which doesn't make it widely applicable. Language models can be evaluated by testing it on documents and texts that the model hasn't seen before and checking the predictive accuracy on those unseen texts. With googles ngram viewer you can search the probability of sequences

of words throughout the web or even only on books because google has uploaded all this data so that you can find the probabilities. All you have to do is write your word sequences separated by a comma and it will compare the proabablities for those sequences on a graph with the probability on the vertical axis and the year on the horizontal to show it's probability over time.