N-grams is the umbrella terminology of the process of processing a certain amount of words at a pace. Through this intake of words, N-grams are capable of building language models by utilizing corpus. A caution of word to all when using N-grams, the sample corpus(ora) will drastically influence the language model. And considering that N-grams are probabilistic models, the trimming and removable of elements are negligible when done correctly.

When it comes to the process of calculating the probabilities of an unigram and bigram, it is important to remember that both require the original text used to create the N-grams. Both N-grams will follow the probability equation P(word1, word2) = P(word1) * P(word2|word1) in which we are multiplying the fraction occurrence of our targeted word, word1, with the fraction product between the total occurrence of word1 & word2 side by side with the sum count of word1 from the raw source text. Thus, emphasizing the importance of utilizing the source text with it comes to calculating the probability.

Of course, when dealing with probability we will sometime come across the number zero which can break the calculation percentage. This is where smoothing comes into play as smoothing does its best to fill over any zero values we might encounter by replacing the zeros with ones. This would of course mislead the total N-gram count be a wide margin; however, smoothing process corrects this by adding in the vocabulary count to the denominator in P(word1) = Count(word1) / total Unigram Num

Knowing how N-grams are calculated and processed is only one half of utilizing the model. Understanding and evaluating the results is critical when utilizing such language processing. The main idea when it comes to laplace smoothing is you want to have your probability approach zero as much as possible. That way, your chances of being accurate reflect with the smoothing probability. By approaching the limitations, your applications in language translation, speech prediction, grammar editing, and speech recognition are just as accurate.