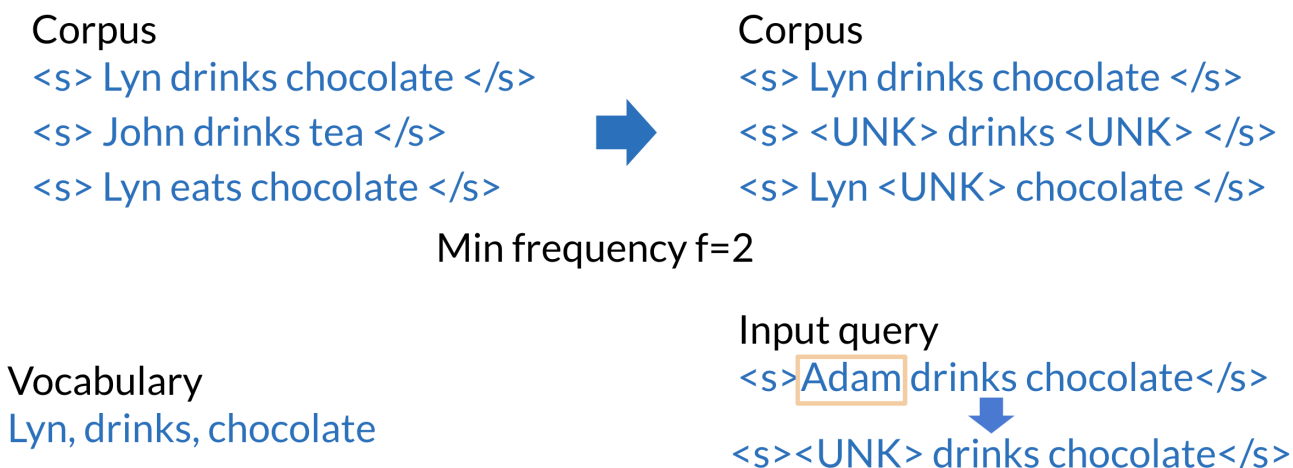# Out of Vocabulary Words

Many times, you will be dealing with unknown words in the corpus. So how do you choose your vocabulary? What is a vocabulary?

A vocabulary is a set of unique words supported by your language model. In some tasks like speech recognition or question answering, you will encounter and generate words only from a fixed set of words. Hence, a **closed vocabulary**.

Open vocabulary means that you may encounter words from outside the vocabulary, like a name of a new city in the training set. Here is one recipe that would allow you to handle unknown words.

- Create vocabulary V

- Replace any word in corpus and not in V by <UNK>

- Count the probabilities with <UNK> as with any other word

Corpus
<s> Lyn drinks chocolate </s>
<s> John drinks tea </s>
<s> Lyn eats chocolate </s>

Corpus
<s> Lyn drinks chocolate </s>
<s> <UNK> drinks <UNK> </s>
<s> Lyn <UNK> chocolate </s>

Min frequency f=2

Vocabulary
Lyn, drinks, chocolate

Input query
<s>Adam drinks chocolate</s>

<s><UNK> drinks chocolate</s>

The example above shows how you can use *min_frequency* and replace all the words that show up fewer times than *min_frequency* by UNK. You can then treat UNK as a regular word.

Criteria to create the vocabulary

- Min word frequency f

- Max |V|, include words by frequency

- Use <UNK> sparingly (Why?)

- Perplexity - only compare LMs with the same V