Creating a Lexicon

Step 1

From the dawn of written language, mankind has made efforts to keep a record of all words in a given language. We should all hopefully be familiar with "lexicons" (*lexicon* is a fancy word for "a list of words") and "dictionaries" (a list of words with their definitions).

The oldest known dictionaries are thought to have been created in roughly 2300 BCE. These dictionaries were Akkadian Empire cuneiform tablets with translations of words between the Sumerian and Akkadian languages. Ancient dictionaries have been found for many languages around the world, including the Chinese, Greek, Sanskrit, Hindvi, and Arabic languages.

The earliest dictionaries in the English language were glossaries of French, Spanish, or Latin words along with definitions of the foreign words in English. The English word *dictionary* was invented by John of Garland in 1220 when he wrote a book, *Dictionarius*, to help with Latin *diction*. Throughout the centuries, numerous other English dictionaries were created; however, it wasn't until Samuel Johnson wrote *A Dictionary of the English Language* in 1755 that a reliable English dictionary was deemed to have been produced.

Now, in the 21st century, dictionaries stored as printed text are becoming less and less popular. When asked about printed dictionaries, a pre-teen cousin of one of the authors had never even used one in the entirety of his life. Instead, online dictionaries, such as those used by services like Google Search or your phone's autocorrect feature, have become the tool of choice. These digital representations of dictionaries are typically able to perform lookups extremely quickly: after searching for a word on Google Search, its definition (and synonyms, and etymology, and statistics about its usage, etc.) are pulled up in approximately half a second.

Figure: The Sumerian–Akkadian dictionary from the Mesopotamian city of Elba (now part of Syria) from around 2300 BCE

Step 2

As computer scientists, we can think of the **lexicon** as an **Abstract Data Type** defined by the following functions:

- **`find(word)`:** Find `word` in the lexicon
- **`insert(word)`:** Insert `word` into the lexicon
- **`remove(word)`:** Remove `word` from the lexicon

These three functions should hopefully sound annoyingly familiar by now, assuming you read the other chapters of the text! We have defined an **Abstract Data Type**, and we are now tasked with choosing a **Data Structure** to use to implement it. In this chapter, we will discuss various possible implementation approaches, focusing on their respective pros and cons.

Throughout this chapter, because languages remain largely unchanging, we will assume that "find" operations are significantly more frequent than both "insert" and "remove" operations. Also, for the same reason, we will assume that we know the size of the lexicon (i.e., the number of words we will be putting into it) before its initial construction, which is quite unlike the applications we've dealt with in the past.