**1. Setting Up the Data for Clustering**

The general goal is to assign a vector to each text that reflects how the diagnostic signs are written in it. There are two ways to differentiate sign usage within a text:

1. Which orthographies are used? This is where the sign modification information is useful to determine which sign orthographies are being used in any given text.
2. Which sign is preferred to write a syllabic value? Due to the homophony of the cuneiform writing system, one syllable can be written with many signs. For example, 'li' can be written with the LI-sign but also with the NI-sign, in which case it would be transliterated as li2

Other variables can be applied to a text as attributes in its vector. (What are these? We talked about things like provenience, city information, scribe information. Also, if we apply different types of variables how can we use a clustering algorithm to treat these vector components as a different entity?).

This section therefore contains two subsections. One groups the diagnostic signs with variants per text. The other discovers the homophonous signs used throughout the corpus and groups different usages per text

## 2. Clustering

### 2.1 Tf-Idf

We want to gather all of the signs with their variant orthographies into each text file and generate a vector which will contain a '1' if the text contains an orthography or sign syllable and a '0' if it does not.

(Here's where I am unsure of how to form the Tf-Idf matrix. The code currently counts ALL usages of the signs per file. But do we want to...

1. Only give each sign usage a 1 or 0 value?
2. Keep the totals and normalize the vectors?

Another thing to consider is if we want to combine Tf-Idf matrices, i.e. the sign orthography variants and the sign syllable variants. Would this be useful? Would we need to distinguish these qualities in the vectors somehow? The same issue occurs if we tack on other variables to the text)

### 2.2 K-Means Clustering

(K-means requires a determination of how many clusters to use. What is this number? How should we determine it?)