Table 1: Data Set Statistics

Name	**Size**	**Query Stats**				
		Mean	**Min**	**Max**	**Median**	**Mode**
USHMM Spoken Names	249	9.80	5	28	9	8
USHMM Names	656	63.77	7	164	54	60
Medical	9,883	10.58	2	30	10	10
US Census Sir Names	88,799	7.83	3	14	8	7

1 Evaluation

1.1 Data Sets

We consider four data sets in our evaluation.

1. USHMM Names
2. USHMM Spoken Names
3. Medical Words
4. US Census Sir Names

Use of these data sets allow us to evaluate Segments over a variety of potential user inputs and environments. For example, the US Census Sir Names data set, which contains roughly 90,000 names, is a good evaluation of the scalability of Segments. Inversely, the USHMM Names data set is an evaluation over a small data set with many languages. A detailed description of each data set along with justification for its use follows.

1.1.1 USHMM Names

1.1.2 USHMM Spoken Names