Type system: As before, **NamedEntityAnnotation** contains the positions of the named entities, as well as the named entities themselves.

ShittySentenceID is a shitty container for a sentence ID for each CAS.

I'm using different models of Lingpipe to do chunking. The collection reader reads the input, the JaCas annotator annotates each sentence using GeneTag, the Fucked annotator annotates each sentence using the Genia model, and the CAS consumer guarantees unique output for gene names from both annotators while only selecting for gene names of a certain confidence and under a certain length. The F1 score is 0.78 so it seems this approach works pretty well.

There really isn't much to say. It's a pretty simple pipeline.