ISMLA 17/18 Exercise 01:
A Greedy Tokenizer For Chinese

1 Project Setup

- Set up a Maven/Java Eclipse project with the usual UimaJ dependencies.
- Save `cmnengdict.jar` somewhere and add it to the build path.
- Add a src/main/resources directory and add it to the project build path.

2 Type System Setup

1. Add a type system descriptor in src/main/resources.
2. Create an annotation type `TokenCMN` with string-valued features `pinyin`, `cat` and `gloss` in addition to the inherited indices.
3. Generate Java code from the XML type system.

3 Analysis Engine Setup and Implementation

1. Write an AnalysisEngine that creates `TokenCMN` annotations based on greedy lookup of document chunks in a `CmnEngDictionary` instance. This instance should be created during initialization.
2. Fill the fields of your `TokenCMN` objects with the information you retrieved from the dictionary in the shape of `SimpleDictionaryEntry` objects.
3. For debugging (and learning some things about Chinese), let your greedy tokenizer print out each `TokenCMN` to a new line in standard output.
4. Test your component on the five Wikipedia snippets in zh-wiki-snippets.tar.gz.
5. Come to one of us **as a group** to show and explain your solution:
    - Johannes' office on Thu, November 9, 1-2 pm
    - Björn's office on Fri, November 10, 2-3 pm
    - after the lecture on Mon, November 13
    - by appointment **before** the next exercise session (November 15)