



# **cTAKES for developers:**

## **How to download and install cTAKES**

### **How to build cTAKES dictionary**

Giaping Zheng  
Children's Hospital Boston



# Installing cTAKES – binary release

1. Java JDK 1.5
  2. Optionally Eclipse IDE
  3. Apache UIMA (with Eclipse plugin)
  4. Extract the binary PEAR files and install each in the following order with runPearInstaller.sh:
    - core, document preprocessor, POS tagger, chunker, context dependent tokenizer, dictionary lookup, LVG, dependency parser (optional), NE contexts, clinical documents pipeline, PAD term spotter (optional), Drug NER (optional)
- Test run:
    - `cvd.sh -desc '/path/to/clinical documents pipeline/clinical documents pipeline_pear.xml'`
  - Installation guide hosted at [SourceForge](#)



# Installing cTAKES – building from source

1. Java JDK 1.5
  2. Eclipse IDE
  3. Apache UIMA (with Eclipse plugin)
  4. Extract the source PEAR packages
  5. Create a new user library and add UIMA JARs
  6. Import and build cTAKES
- **Test Run:**
    - Run "UIMA\_CVD--clinical\_documents\_pipeline"
    - Load '/path/to/clinical\_documents\_pipeline/clinical\_documents\_pipeline\_pear.xml'
  - **Installation guide hosted at [SourceForge](#)**
    - To compile from command line, please read the online doc



# Installing cTAKES – configuring dictionaries

- cTAKES is only shipped with sample dictionaries
- UMLS is required to make real use of cTAKES
- Enable SQL dictionary
  - Download SQL Wrapper class [files](#) by Daniel Armbrust
  - Uncomment  
`edu.mayo.bmi.uima.core.resource.JdbcConnectionResourceImpl` and  
`import WrappedConnection`



## Installing cTAKES – configuring dictionaries

- Acquire a license (free of charge) from NLM
- Building UMLS dictionary for disease/disorder, sign/symptom, procedure, and anatomical site terms
  - Subset the full UMLS metathesaurus using MetaMorphoSys
  - Tokenize the terms with cTAKES tokenizer
  - Load the dictionary into a MySQL database
  - Configure cTAKES to use the database
- Building RxNORM dictionary for medication terms
  - Subset the full RxNORM
  - Tokenize the terms with cTAKES tokenizer
  - Generate Lucene index
  - Configure cTAKES to use the index
- Helper script ([dictgen](#)) is available on the forum