

cTAKES 3.2 User Install Guide

Contents of this Page

- [Prerequisites](#)
- [Install cTAKES](#)
- [\(Recommended\) Add UMLS access rights](#)
- [Process documents using cTAKES](#)
 - [CAS Visual Debugger \(CVD\)](#)
 - [Collection Processing Engine \(CPE\)](#)
- [Next Steps](#)

cTAKES 3.2 Links

[Apache cTAKES website](#) (including [downloads](#))

Documentation:

- A [pamphlet/manual](#) on cTAKES basics
- [cTAKES 3.2 Main Wiki page](#)
- [cTAKES 3.2 User Install Guide](#)
- [cTAKES 3.2 Developer Install Guide](#)
- [cTAKES 3.2 Component Use Guide](#)
- [cTAKES 3.2 Dictionaries and Models](#)
- [Documentation Conventions](#)

These instructions are for end users. With these instructions you can install Apache cTAKES, configure it, and use it to process text (typically text associated with a medical record). If you were planning to expand, change, or modify the code within cTAKES, refer to the [cTAKES 3.2 Developer Install Guide](#).

These instructions will cover installation and a test of the main product including trained models for sentence detection and tagging parts of speech, dictionaries from a subset of the UMLS, the LVG resource, etc. Optional components are described in the [Component Use Guide](#).

Once you have finished installing cTAKES and its separately-bundled resources, you will be able to see what cTAKES is capable of.

Prerequisites

| Step | Example |
|--|---|
| <p>1. Make sure you have Java 1.7 or higher. Most systems come with Java already installed.</p> <p>Run this command to check your version.</p> <p>Windows and Linux:</p> <pre>java -version</pre> | <p>Windows:</p> <pre>C:\>java -version java version "1.7.0_20" Java(TM) SE Runtime Environment (build 1.7.0_20-b02) Java HotSpot(TM) Client VM (build 16.3-b01, mixed mode, sharing)</pre> <p>Linux:</p> <pre>tbleeker@system:/\$ java -version java version "1.7.0_22" OpenJDK Runtime Environment (IcedTea6 1.10.1) (6b22-1.10.1-0ubuntu1) OpenJDK 64-Bit Server VM (build 20.0-b11, mixed mode)</pre> |

Install cTAKES

| Step | Example |
|------|---------|
|------|---------|

1. Navigate to the cTAKES [downloads page](#) on the Apache site and download the **binary** package. Select a mirror site and press the Change button to modify the URL to your desired mirror location before doing the download or accept the default.

Windows:

Download the ZIP file.

Linux:

Use wget to obtain the *.TAR.GZ file.
wget <URL of the file from downloads>



The download time will be commensurate with ~165MB of data.

Windows:

| File | Signatures |
|--|--|
| apache-ctakes-3.1.0-bin.tar.gz | md5 sha1 asc |
| apache-ctakes-3.1.0-bin.zip | md5 sha1 asc |
| apache-ctakes-3.1.0-src.tar.gz | md5 sha1 asc |
| apache-ctakes-3.1.0-src.zip | md5 sha1 asc |

Linux:

```
HTTP request sent, awaiting response... 200 OK
Length: 763500777 (728M) [application/x-gzip]
Saving to: `apache-ctakes-3.2.1-bin.tar.gz'
```

```
13% [=====>] 106,548,331 1.13M/s
eta 11m 9s
```

2. (Optional but recommended) [Verify the downloaded files](#) against a file signature to ensure you have the proper and complete file.

No example

3. Unzip the file you downloaded into a directory that you want to be the cTAKES install location. The compressed files contain a single directory at the top level. This folder we will call <cTAKES_HOME>. You will need to refer to this directory later.

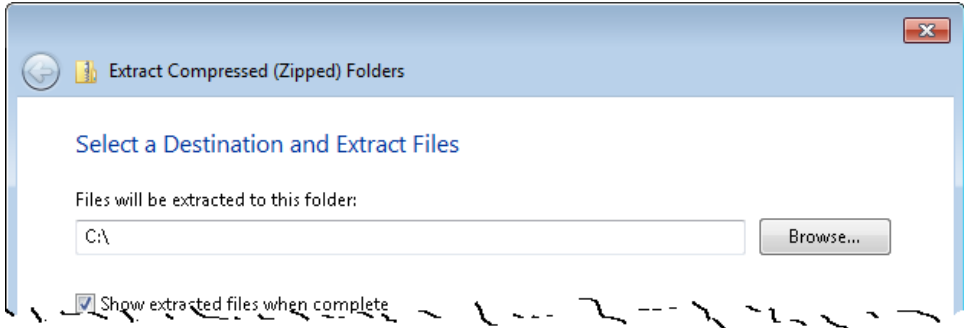
Windows:

C:\apache-ctakes-3.2.1

Linux:

/usr/local/apache-ctakes-3.2.1

Windows:



Linux:

```
tar -xvf apache-ctakes-3.2.1-bin.tar.gz -C /usr/local
```

4. Download the cTAKES resources ZIP file with a matching version from the ctakesresources project ([More information on cTAKES models](#)). These resources are required to operate cTAKES.

i Due to licensing considerations resources are hosted at an external location. For ease of installation, a single package was created with all the resources you will need. Licensing for these resources is found within the download.

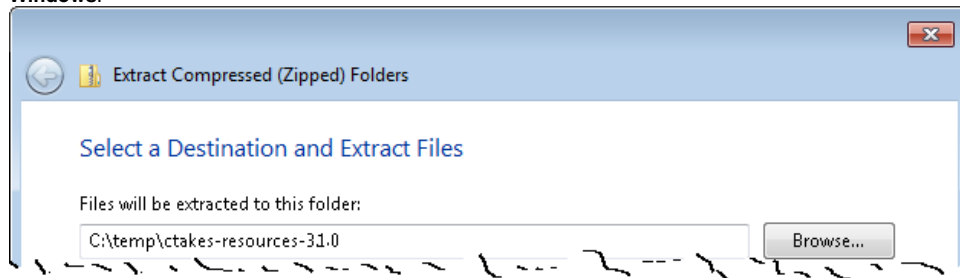
i Download time will be commensurate with 1GB of data.

Unzip the cTAKES resources file into a temporary location.

5. Copy (or move) the resources to cTAKES_HOME.
Copy the contents of the temporary resources directory (and all sub-directories) to <cTAKES_HOME>/resources.

i There may be conflicts while taking this action. Overwrite the cTAKES_HOME files with those in the resources download.

Windows:



Linux:

```
cd /tmp
wget http://sourceforge.net/projects/ctakesresources/files/ctakes-resources-3.2.1.zip
sudo unzip ctakes-resources-3.2.1.zip
```

Windows:

```
xcopy /s C:\temp\ctakes-resources-3.2.1\resources C:\apache-ctakes-3.2.1\resources
```

Linux:

```
cp -R /tmp/resources/* /usr/local/apache-ctakes-3.2.1/resources
```

Mac OSX:

```
ditto /tmp/resources/* /usr/local/apache-ctakes-3.2.1/resources
```

(Recommended) Add UMLS access rights

! In the initial setup cTAKES will recognize only few sample concepts in text. If you wish to perform named entity recognition or concept identification for anything other than these few words, you will need to 1) obtain the rights to use UMLS resources 2) add those credentials to cTAKES, and 3) use an aggregate that makes use of those UMLS resources. If you don't, cTAKES will work but won't recognize much.

| Step | Example |
|--|------------|
| 1. If you do not have a UMLS username and password, you may request one at UMLS Terminology Services . | No example |

2. Edit the following files. Find the line in each script that runs java and add the ctakes.umluser and ctakes.umlspw parameters to the java command with your credentials. Make sure you substitute your actual ID and password if you cut and paste the example.

Windows:

```
<CTAKES_HOME>\bin\runtakesCVD.bat
<CTAKES_HOME>\bin\runtakesCPE.bat
```

Linux:

```
<CTAKES_HOME>/bin/runtakesCVD.sh
<CTAKES_HOME>/bin/runtakesCPE.sh
```

```
java -Dctakes.umluser=<YOUR_UMLS_ID_HERE> -
Dctakes.umlspw=<YOUR_UMLS_PASSWORD_HERE> -
cp ...
```

For example, if your username and password were literally myusername and mypassword, you could insert them before the -cp option so the start of the java command would look like this:

```
java -Dctakes.umluser=myusername -
Dctakes.umlspw=mypassword -cp ...
```


Process documents using cTAKES

This version allows you to test most components bundled in cTAKES in two different ways:

1. Using the bundled UIMA CAS Visual Debugger (CVD) to view the results stored as XCAS files or run the annotators
2. Using the bundled UIMA Collection Processing Engine (CPE) to process documents in cTAKES_HOME/testdata directory

You will need a windowing environment on Linux to run these tools.

CAS Visual Debugger (CVD)

| Step | Example |
|---|---|
| <p>1. Open a command prompt and change to the cTAKES_HOME directory.</p> <div>  It is best if <CTAKES_HOME> is your current directory. The scripts will change directories, so being home to run the command is best. </div> | <p>Windows:</p> <pre>cd \apache-ctakes-3.2.1</pre> <p>Linux:</p> <pre>cd /usr/local/apache-ctakes-3.2.1</pre> |
| <p>2. Start the CAS Visual Debugger by running this command: The application may take a minute to start on slower hardware.</p> | <p>Windows:</p> <pre>bin\runtakesCVD.bat</pre> <p>Linux:</p> <pre>bin/runtakesCVD.sh</pre> |

3. Copy the example text from the next cell in this table and paste the contents into the Text section of CVD, replacing the text that is already there.

You can also download a copy of the file from [here](#)

Dr. Nutritious

Medical Nutrition Therapy for Hyperlipidemia

Referral from: Julie Tester, RD, LD, CNSD

Phone contact: (555) 555-1212

Height: 144 cm Current Weight: 45 kg Date of current weight: 02-29-2001

Admit Weight: 53 kg BMI: 18 kg/m2

Diet: General

Daily Calorie needs (kcal): 1500 calories, assessed as HB + 20% for activity.

Daily Protein needs: 40 grams, assessed as 1.0 g/kg.

Pt has been on a 3-day calorie count and has had an average intake of 1100 calories.

She was instructed to drink 2-3 cans of liquid supplement to help promote weight gain.

She agrees with the plan and has my number for further assessment. May want a Resting

Metabolic Rate as well. She takes an aspirin a day for knee pain.

4. An analysis engine (AE) needs to be loaded in order to process text. If you installed the UMLS resources, use

```
AggregatePlaintextFastUMLSP  
rocessor.xml
```

in this step.

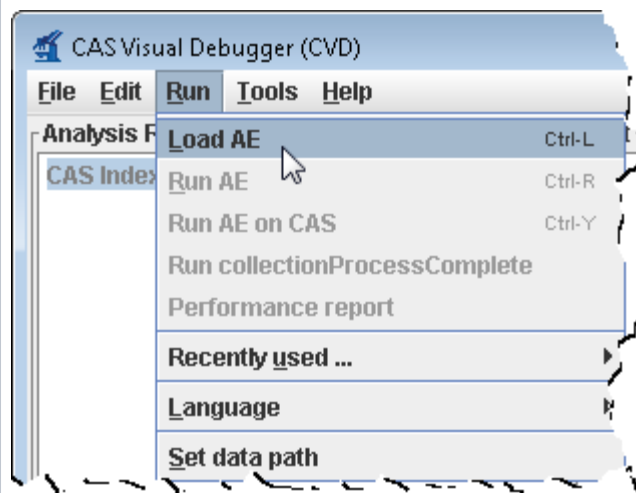
Use the **Run-> Load AE** menu bar command. Navigate to the file

```
<CTAKES_HOME>  
/desc  
/ctakes-clinical-  
pipeline  
/desc  
/analysis_engine  
  
/AggregatePlaintextFastUMLSP  
rocessor.xml
```

Click **Open**.

Loading the analysis engine may take a minute. Once the analysis engine has successfully loaded you should see a tree in the **Analysis Results frame**:

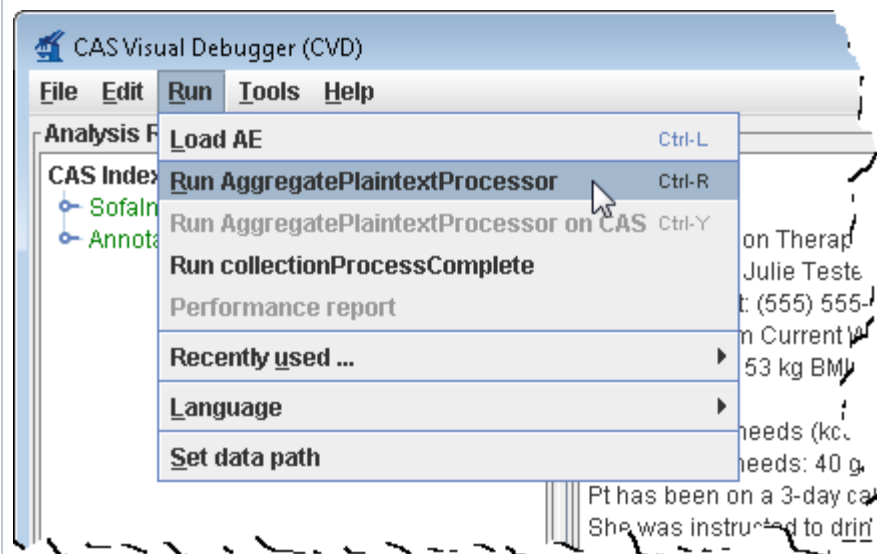
```
CAS Index  
Repository  
* SofaIndex [0]  
* AnnotationIndex  
[1]
```



5. From the menu bar, click **Run -> Run AggregatePlaintextFastUMLSPProcess** or.

Note: If you would like to TEST some simple annotators to ensure it's working without UMLS, you can just load:

/desc/ctakes-core/desc/analysis_engine
/SentencesAndTokensAggregate.xml



6. You'll get a list of all the annotations for this clinical document in the Analysis Results frame. Annotations such as named entities, division by sentence, etc from the pipeline are viewable. To see one, in the **Analysis Results** frame, click on the key in front of:

CAS Index
Repository
* AnnotationIndex
* uima.tcas.
Annotation
* org.apache.
ctakes.tpsystem.
type.textsem.
IdentifiedAnnotation
n
* org.apache.
ctakes.tpsystem.
type.textsem.
EventMention

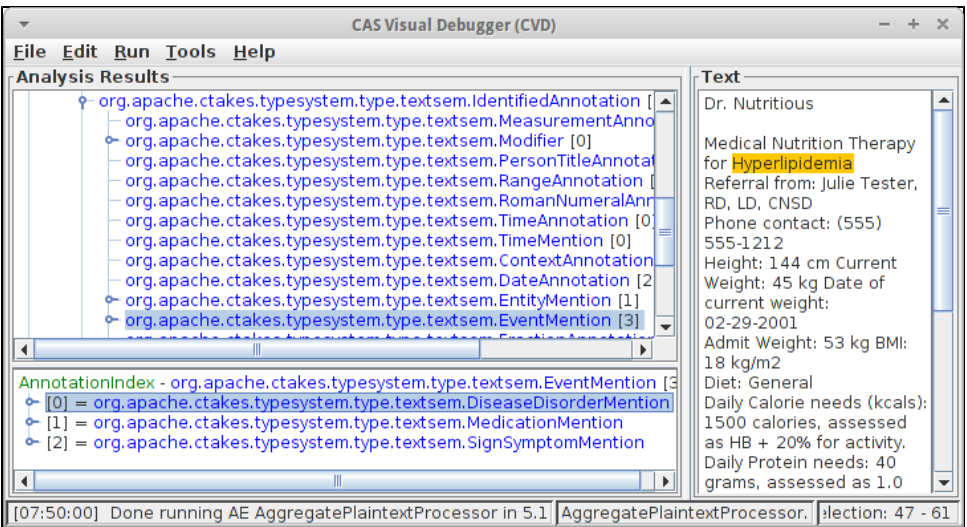
This will show an AnnotationIndex in the lower frame. Select any annotation in that lower frame and you will see the text discovered in the text frame on the right like the concept of the disease/disorder Hyperlipidemia.

For a medication example select this

CAS Index
Repository
* AnnotationIndex
* uima.tcas.
Annotation
* org.apache.
ctakes.tpsystem.
type.textsem.
IdentifiedAnnotation
n
* org.apache.
ctakes.tpsystem.
type.textsem.
EventMention
* org.apache.
ctakes.tpsystem.
type.textsem.
MedicationMention

Now select items in the lower frame to see the text being annotated.

You may close the **CAS Visual Debugger (CVD)** application if you wish.



Collection Processing Engine (CPE)

| Step | Example |
|------|---------|
|------|---------|

1. Open a command prompt and change to the cTAKES_HOME directory:



It is best if <cTAKES_HOME> is your current directory. The scripts will change directories, so being home to run the command is best.

Windows:

```
cd \apache-ctakes-3.2.1
```

Linux:

```
cd /usr/local/apache-ctakes-3.2.1
```

2. Create a directory for some test data.

Windows: mkdir testdata

3. Download this [sample file](#) and place it into the testdata directory.

No example

4. Start the collection processing engine by running this command:
The application may take a minute to start on slower hardware.

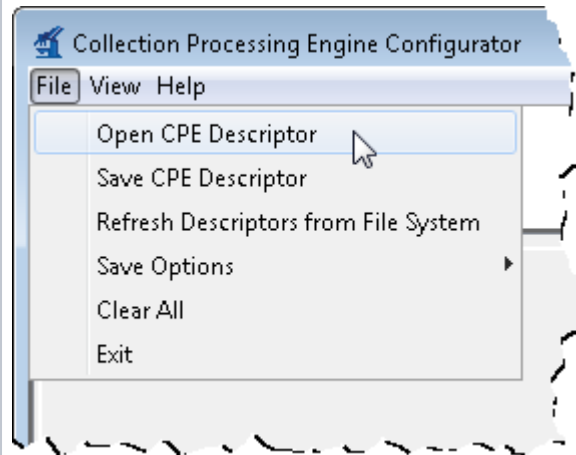
Windows:

```
bin\runtakesCPE.bat
```

Linux:

```
bin/runtakesCPE.sh
```

5. This will bring up the Collection Processing Engine Configurator. In the Menu bar click **File > Open CPE Descriptor**



6. Navigate to the following file, which uses the AggregateCdaProcessor

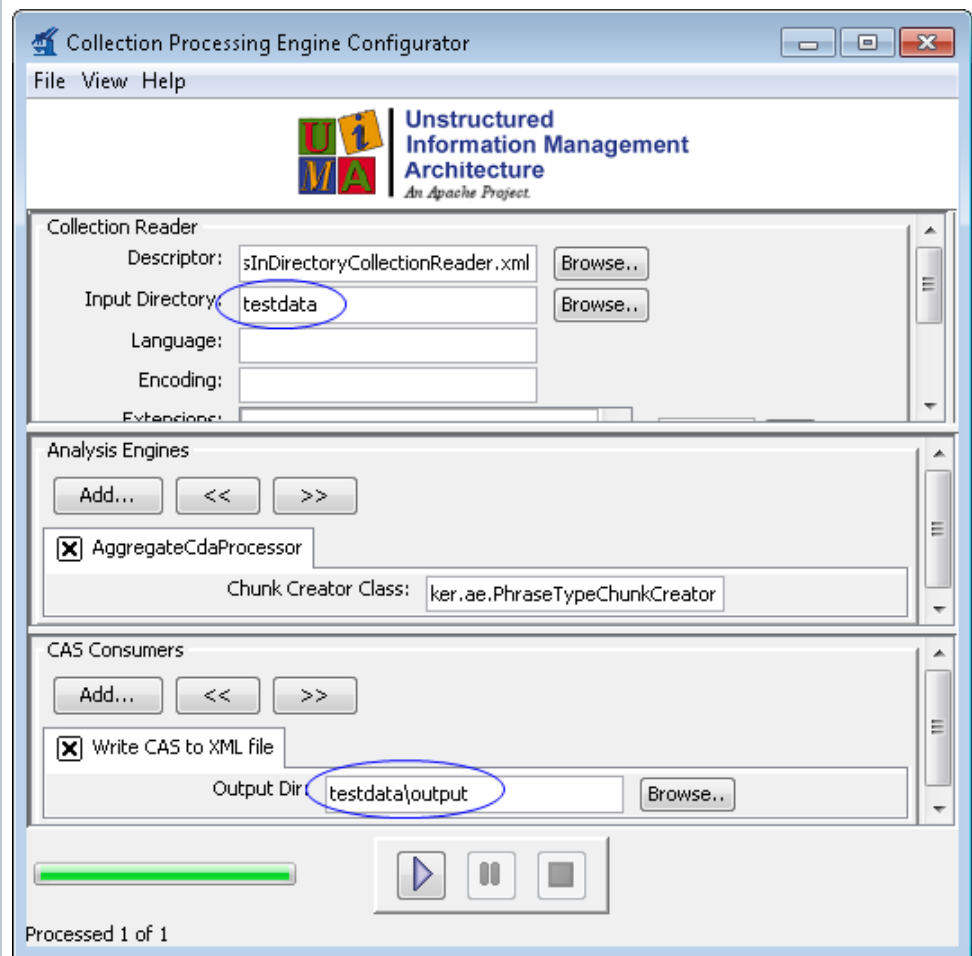
```
<cTAKES_HOME>
  /desc
    /ctakes-clinical-
pipeline
  /desc

/collection_processing_engin
e
  /test1.xml
```

Click **Open**.

No example

7. Change the Collection reader input directory to testdata, which contains a CDA file(s).
Within the CAS Consumers pane of the same window, change the output directory to testdata/output




8. Click the Play button (green/blue play arrow near the bottom).

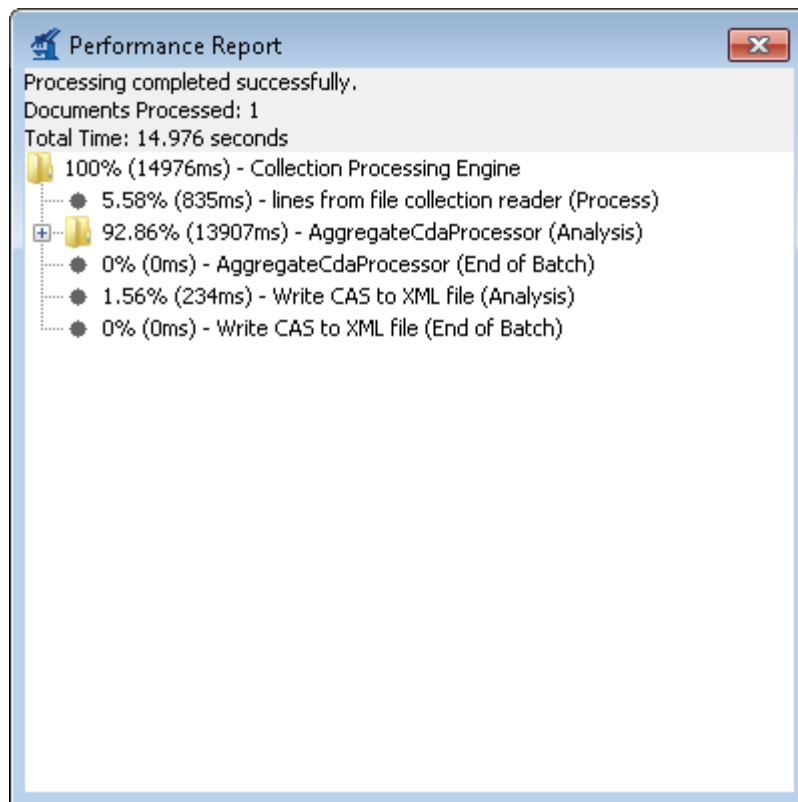


What just happened? You placed a sample CDA document into the input of a pipeline. The pipeline was a file system reader that will process all files in a directory. The processing was accomplished by the chunker cTAKES component (noted by the name of the Analysis Engine pictured). One resulting file for each input file was placed into the output directory. This output file is an XML file that annotates the noun phrases and verb phrases.



9. You should see that one document was processed. You did process a collection of documents. In this case the collection only contained one just to show how to do it. Close the results window.

 This example of using the CPE GUI did not use the UMLS resources. If you wish to perform named entity recognition or concept identification for anything other than a few words, you will need to 1) obtain the rights to use UMLS resources 2) add those credentials to cTAKES, and 3) use an aggregate that makes use of those UMLS resources (see above).



10. Close the CPE application. You may be prompted to save changes. Since this was just a test you may click the **No** button.

No example

Using the same CVD and CPE programs in the manner described above, you can test all the other components. The analysis engines and collection processing engines shipped with cTAKES for some of the annotators are described in the following table.

 cTAKES 3.1 binary distributions did not include test data. Loading the CPE descriptors into the CPE tool will require resetting the input and output directories. Test files could be obtained from the [cTAKES 2.5 release binary distribution](#). Look for a testdata directory in cTAKES_HOME.

| Annotator | Description | Example Aggregate Analysis Engine (AE) | Example Collection processing Engine (CPE) |
|----------------------------|--|---|---|
| Clinical Document Pipeline | The complete cTAKES pipeline to obtain majority of cTAKES annotations | <cTAKES_HOME>/desc/ctakes-clinical-pipeline/desc/analysis_engine/AggregatePlaintextUMLSPprocessor.xml | <cTAKES_HOME>/desc/ctakes-clinical-pipeline/desc/collection_processing_engine/test1.xml |
| Chunker | Obtain cTAKES chunk annotations | NA | NA |
| Dependency Parser | Obtain dependency parsing tree | <cTAKES_HOME>/desc/ctakes-dependency-parser/desc/analysis_engine/ClearParserSRLTokenizedInfPosAggregate.xml | <cTAKES_HOME>/desc/ctakes-dependency-parser/desc/collection_processing_engine/ClearParserTestCPE.xml |
| Drug NER | The annotator to obtain drug annotations | <cTAKES_HOME>/desc/ctakes-drug-ner/desc/analysis_engine/DrugAggregatePlaintextUMLSProcesor.xml | <cTAKES_HOME>/desc/ctakes-drug-ner/desc/collection_processing_engine/DrugNER_PlainText_CPE.xml |
| Dictionary Lookup | Mapping cTAKES annotations to dictionaries (e.g., SNOMED_CT or RxNorm) | <cTAKES_HOME>/desc/ctakes-dictionary-lookup/desc/analysis_engine/TestAggregateTAE.xml | NA |
| Relation Extractor | Annotate certain relations between certain Event, Entity, and Modifier annotations | <cTAKES_HOME>/desc/ctakes-relation-extractor/desc/analysis_engine/RelationExtractorAggregate.xml | N/A |
| Smoking Status | The annotator to obtain document or patient-level smoking status | <cTAKES_HOME>/desc/ctakes-smoking-status/desc/analysis_engine/SimulatedProdSmokingTAE.xml | <cTAKES_HOME>/desc/ctakes-smoking-status/desc/collection_processing_engine/Sample_SmokingStatus_output_flatfile.xml |
| Side Effect | The annotator to find side effect mentions and sentences from clinical documents | <cTAKES_HOME>/desc/ctakes-side-effect/desc/analysis_engine/SideEffectAggregateTAE_UMLS.xml | <cTAKES_HOME>/desc/ctakes-side-effect/desc/collection_processing_engine/SideEffectCPE.xml |

Next Steps

The [cTAKES 3.1 Component Use Guide](#) will help you to understand, in great detail, each of the cTAKES components that have been installed. In some cases you can learn how to improve the components.

Also, before you go on to process text in production, you will want to consider [dictionaries](#) and [models](#). If you did not obtain the rights yet to the UMLS resources and models, you will want to do so. Be aware, the models have been trained on data that may not match your data well enough to be effective. In some cases you might want to [modify the dictionaries and train models](#) using your own data.