



# Module 1 Session 1

---

## Introduction to GATE Developer





# GATE

- Open-source software framework and set of ready solutions for text/natural language processing
- Re-usable abstractions for documents, format conversion, corpora, annotations, storage, algorithms, ...
- A graphical user interface to interactively develop solutions (GATE GUI, GATE Developer)
- A (Java) library providing a programming API for using the abstractions
- An infrastructure of pluggable components (GATE Plugins)
- Ready-made solutions to get you started
- Companion software for semantic search (Mimir) a.o.
- Scalable from laptop to massive processing on servers
- ...



# About this tutorial

- This tutorial will get you started with the GATE graphical user interface (GUI), also known as “GATE Developer”
- It will be a hands-on session.  
Please try things out in GATE as the topics are presented.  
If there are questions or problems please interrupt, we will be happy to help immediately!
- Things suggested for you to try yourself are in **red**.
- **Start GATE on your computer now (if you haven't already)**
- Please don't jump ahead: if you're already familiar with some topics, perhaps you can help your neighbour if they get stuck.
- This tutorial is about how to **use** the various components.  
Later sessions and modules are also about the underlying functionality.



# GATE GUI

Menu Bar

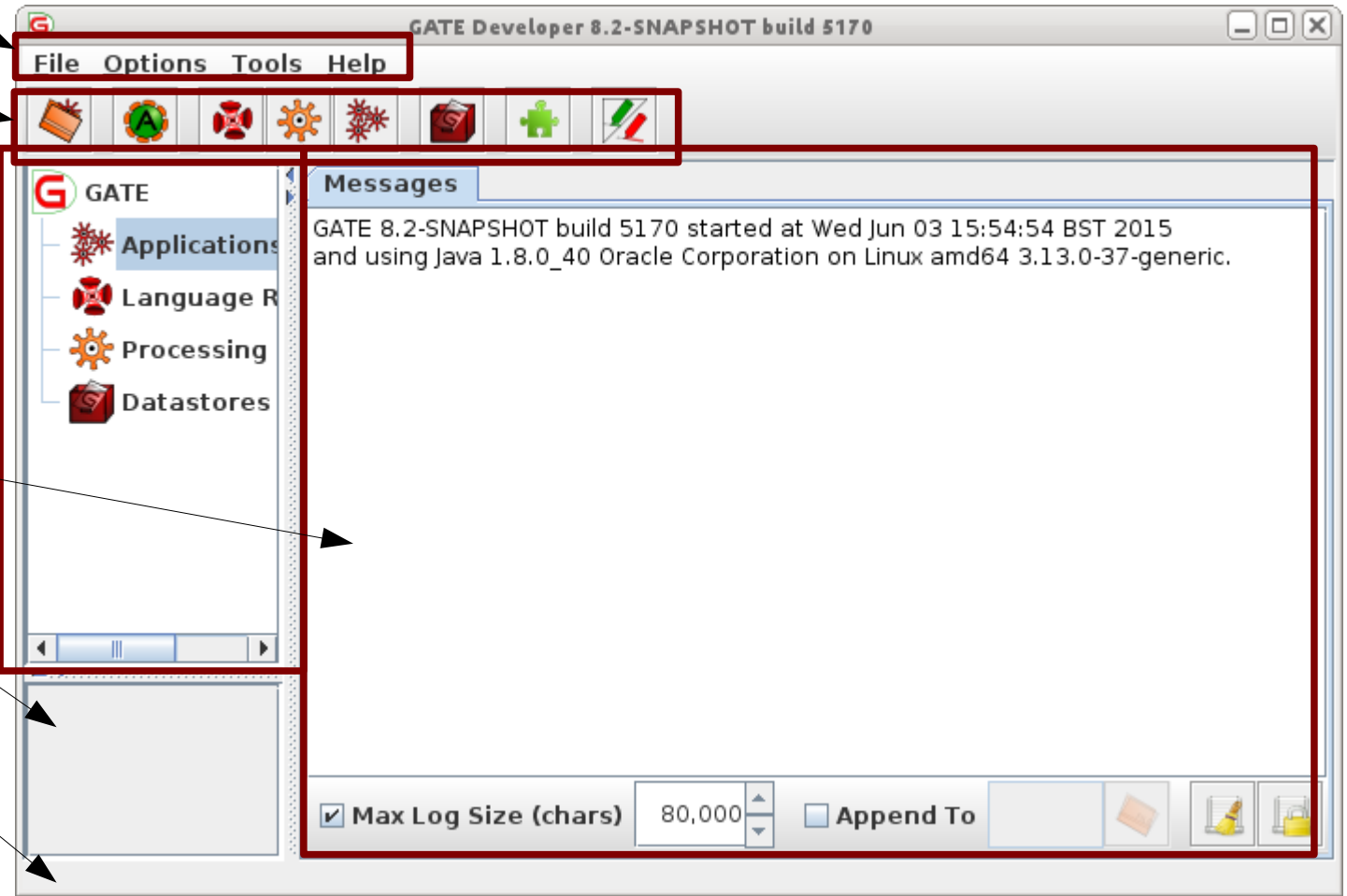
Shortcut Buttons

Resources Pane

Display Pane

Resource Features

Messages





# Resources

- Most “things” you use within GATE are “**resources**”:
- **Language resources** (LRs) are documents, document collections, ontologies ...  
A collection of documents is known as a **corpus**
- **Processing resources** (PRs) are programs that operate on text within the documents, and often create or modify annotations
- **Data stores** are for storing documents and corpora for later use (on your hard disk: directories)
- **Applications** (“pipelines”) are sequences of processing resources that run on one or more documents
- These resources are *listed* in the **Resources Pane**
- These resources can be *viewed* in the **Display Pane**



# Displaying Resources

- When you first open GATE, the display pane will show messages from the system in the “Messages” tab
- The display pane displays whatever elements you are currently working with, e.g. an application, a document or a processing resource, each in its own *tab*
- Double clicking on a resource will display it
- Tabs along the top of the display pane allow to choose which of the open resources to display
- Clicking on a tab displays that view e.g. “Messages” tab shows messages



# Create New Document

- **Right click “Language Resources” → New → GATE Document**
- A dialog for entering or changing parameters appears, ignore this for now
- **Click OK**
- “GATE Document\_<id>” added to “Language Resources”
- **Double click that document name**
- A tab is opened in the display pane, showing the empty document. You can **enter some text** there if you want.



# Empty Document

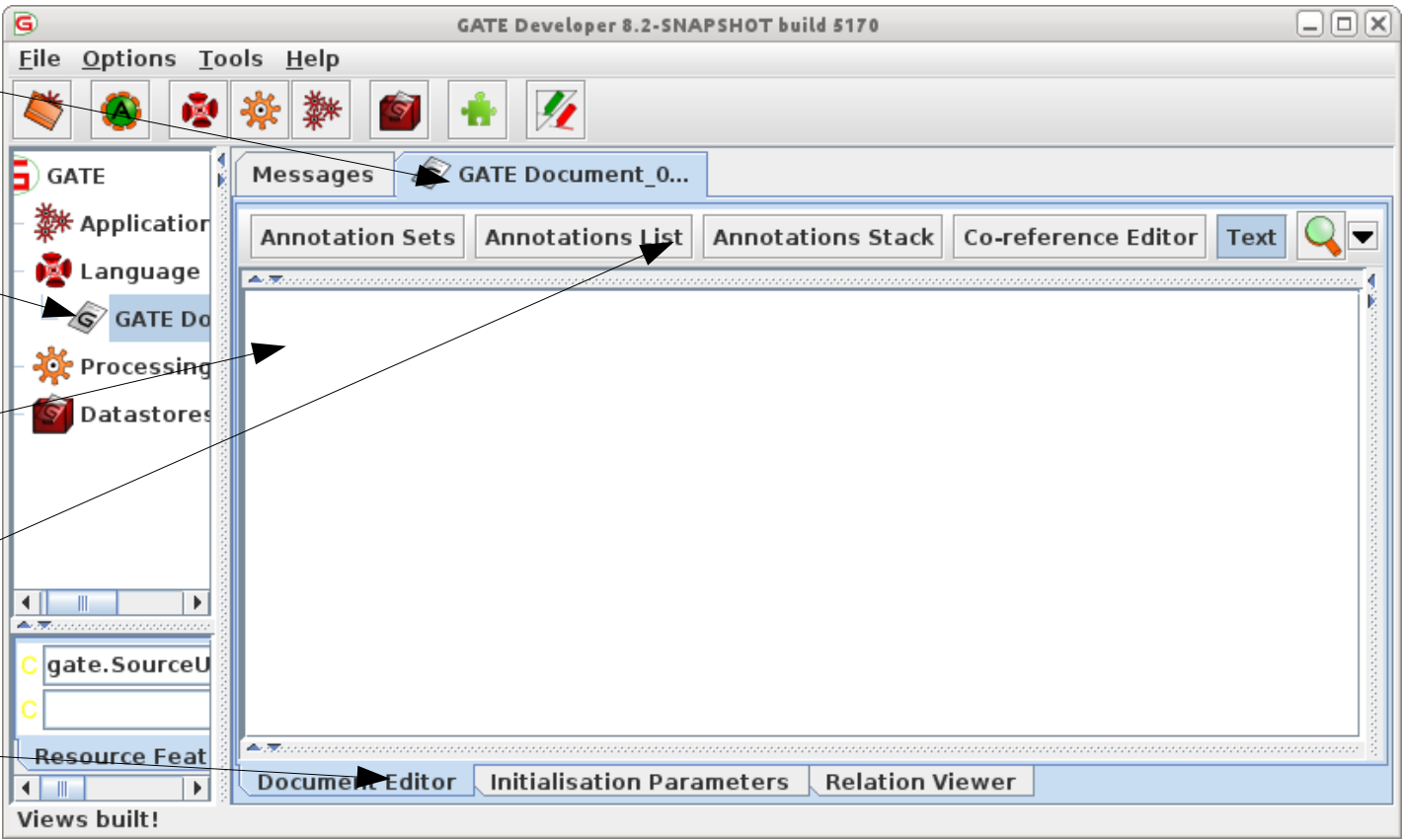
Document Tab

Document Name

Document Editor

Document Editor Buttons

Document Resource Views







# Document Editor

---

- The Document Editor is shown as a new Tab in the Display Pane, alongside the Message Pane
- There are buttons on the top of the Editor, e.g. “Annotation Sets” – we will learn about them later.
- There are tabs at the bottom of the Document Tab: these show different “Views” of the document.
- The small pane in the lower left shows the “document features” (optional information associated with the document resource as key/value pairs)



# Document Editor

- **Right click the document tab → “Hide”**
  - This will close the document tab but not the document
  - **Double click the document name to re-show the tab**
  - The document is shown in a new document tab again
  - **Right click the name → “Close”**
  - This removes the document and the document tab
    - No warning that a resource is not saved!
    - No way to undo a close!
- Also: no warnings about overwriting a file when saving!

# Simple operations on resources



- Right clicking on the name of a resource in the resource pane gives access to a menu of actions
- Double clicking on the name of a resource opens a view of the resource in the display pane (triple clicking the name can be used to rename)
- Selecting a resource instance and pressing the Delete (Mac: Fn+backspace) key will generally close it
- You can also right click and then select “Close”



# Parameters

- Resources can have parameters which need to get specified when the resource is created (initialization): **Initialization Parameters** cannot be changed (these will be called “init parameters” for short)
- Init parameters specify how a resource is created, e.g. the location of a document to load
- Processing resources can also have parameters which can be changed for each run: **Runtime Parameters**
- Runtime parameters configure what a processing resource does, e.g. if some processing is case-sensitive or not.



# Loading a document

---

- GATE can read and load documents in many formats: e.g. plain text, HTML, XML, PDF, Word, CoNLL
- GATE can load documents from files and from URLs
- For formats with markup such as HTML, XML the **markupAware** parameter causes the markup to get converted to annotations (true) or included in the text (false).
- When a document is loaded, it gets converted to GATE internal format as document text + annotations.



# New Document Parameters

Required?

Document Resource Name

Parameter Name

Parameter Value

Parameters for the new GATE Document

Name:

Name	Type	Required	Value
collectRepositioningInfo	Boolean	<input checked="" type="checkbox"/>	false
encoding	String	<input type="checkbox"/>	
markupAware	Boolean	<input checked="" type="checkbox"/>	true
mimeType	String	<input type="checkbox"/>	
preserveOriginalContent	Boolean	<input checked="" type="checkbox"/>	false
sourceUrl	URL	<input checked="" type="checkbox"/>	<input type="text"/>
sourceUrlEndOffset	Long	<input type="checkbox"/>	
sourceUrlStartOffset	Long	<input type="checkbox"/>	

OK Cancel Help

Help (or F1)

File Chooser



# Loading a document

- To load a document:
  - right click on Language Resources → “New → GATE Document”  
OR
  - File menu → New Language Resource → GATE Document
- Use the sourceURL parameter to specify the document to be loaded:
  - type the filename or URL, or
  - click the file browser icon to navigate to the correct document.
- **Load a file from your hands on materials:**  
**annie-hands-on → news-texts → ft-airlines-27-jul-2001.xml**
- **Load a web page – for this the http:// or https:// part of the URL is required, e.g. http://news.bbc.co.uk**
- **Load a PDF file from your hands on materials:**  
**annie-hands-on → example-texts → Convention\_ENG.pdf**

# Initialisation parameters

---

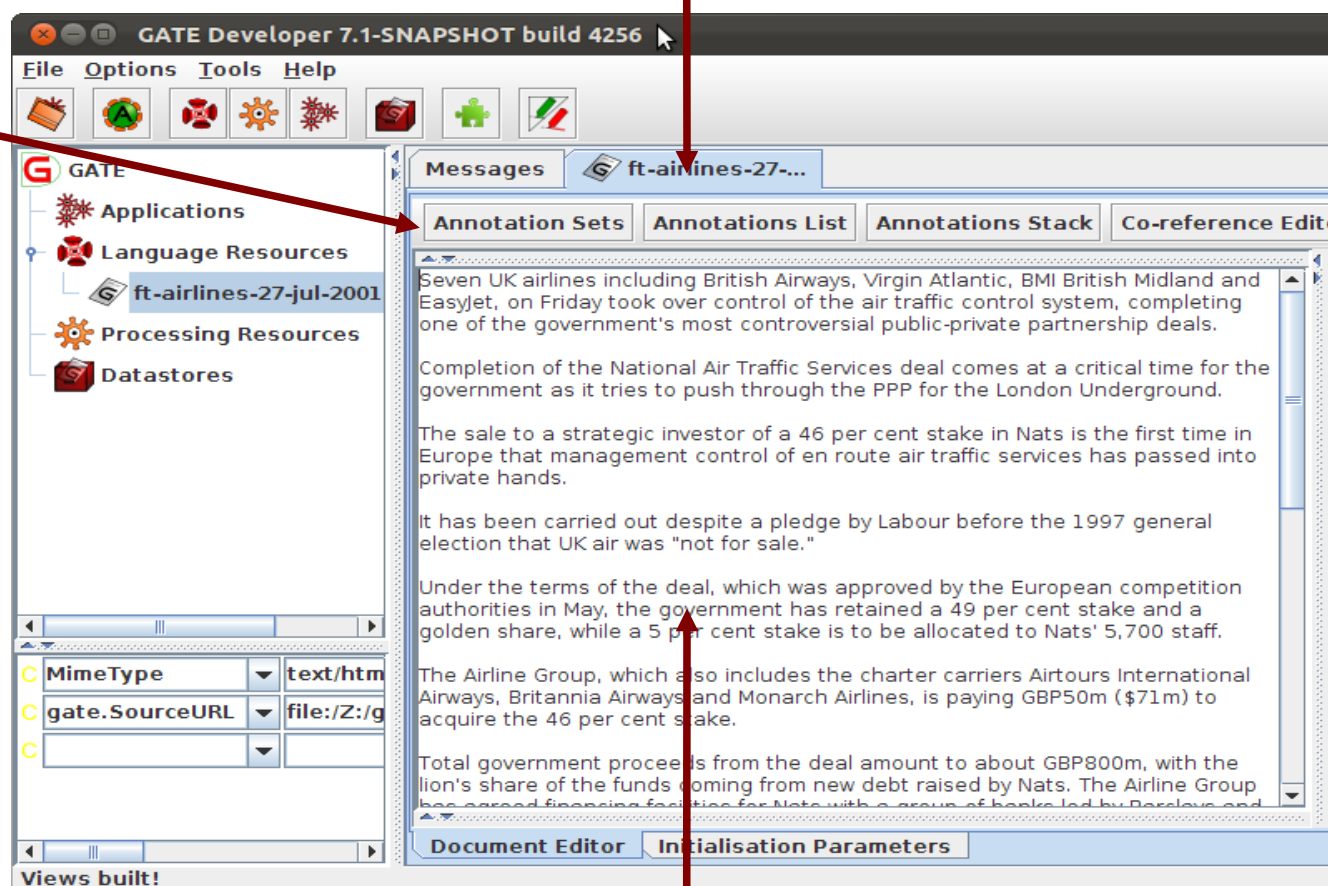
- A document has a variety of init parameters: some compulsory and some optional
- Compulsory parameters have a tick in the “Required” box
- You can provide your own name or use the default name GATE provides (document name + a unique ID, which prevents confusion with multiple copies of the same document)
- Note that the same approach to naming applies with all other resources (such as PRs)



# Document viewer

Highlighted tab is the resource currently being viewed

Document viewer buttons



Document



# Opening and closing documents

- To view a document, double click on the document name in the Resources pane
- To close, right click on the document name and select “Close”
- To hide a document, while leaving it loaded, right click on the document name or the tab and select “Hide”
- The document viewer buttons at the top of each document tab let you select different views
- **Load the “ft-airlines-27-jul-2001.xml” file from annie-hands-on/news-texts**



# Annotations

- Annotations are central to GATE.
- Annotations represent aspects of the text you want to analyze: Words, Sentences, Dates, Person Names
- Annotations are named by their type, e.g. "Person"
- Annotations are grouped into sets
- Annotation consists of
  - Annotation type
  - start and end offsets
  - set of features, each feature is an arbitrary name/value pair, e.g. orth="upperInitial"

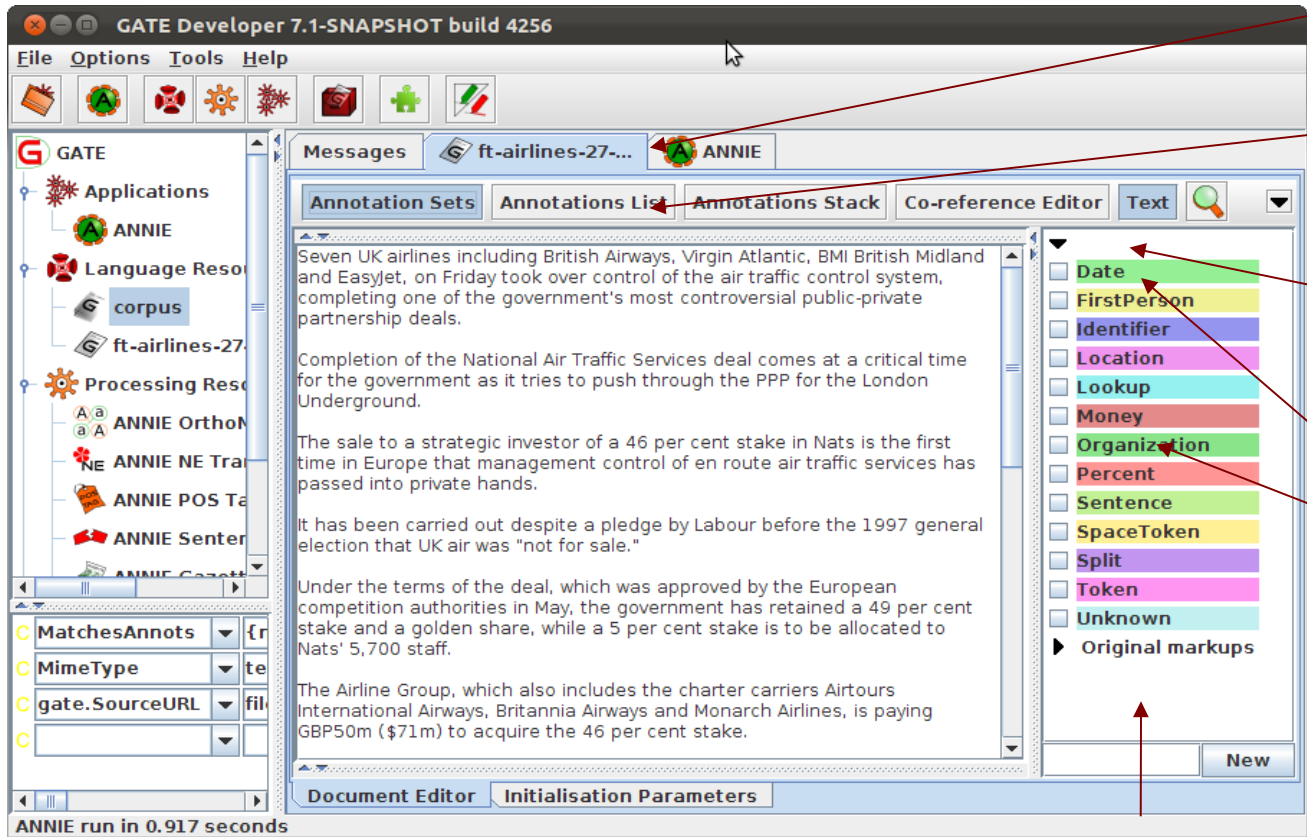


# Annotation Sets

- Annotations are grouped into sets
- Each set can contain any number of annotations of any type
- You can create and organize your annotation sets as you wish.
- Predefined sets:
  - Default set (empty name): cannot be deleted
  - “Original markups”: annotations from the markups in the file
  - “Key”: by convention, used for known correct annotations (→ session 2)
- **Click the “Annotation Sets” button in the document viewer**



# Annotation Sets



Tabs

Document Viewer Buttons

Default annotation set

Annotation types

Original markups annotation set



# Viewing annotations

- Clicking on the Annotation Sets button opens a new pane on the right hand side inside the document view (Annotation Sets view)
- Default (unnamed) set contains some examples of annotations
  - **Click on the ► to display the annotation types belonging to that set**
- You should see types such as Location, Date, Person etc.
  - **Click the check box for an annotation type to view all the annotations of that type in the document**



# A closer look at the annotations

- **Click the Annotations List button from the menu above the Display pane**
- A new pane is created inside the document viewer that shows a table that lists all *selected* annotations.
- Table shows annotation type, annotation set, offsets, annotation id, and features
- **Select a row in the table to highlight the annotation in the text**
- **Click on a column heading to sort according to the header**
- There are also other annotation views possible such as the Annotation Stack and Coreference Editor: we'll look at these later



# Annotations

Date annotation

The screenshot shows the GATE Developer 7.1 interface. The main text area contains several paragraphs with date annotations. A red arrow points from the text "Date annotation" to the "Date" checkbox in the right-hand list of annotation types. Another red arrow points from the "Date" row in the "Annotations table" to the "Date" checkbox in the list.

**Annotations table**

Type	Set	Start	End	Id	Fea
Location		6	8	1273	{locType=country, matches=[1273, 1284
Date		98	104	1278	{kind=date, rule1=GazDate, rule2=Date
Percent		449	460	1255	{rule=PercentBasic}
Location		496	502	1282	{locType=region, rule1=InLoc1, rule2=L
Date		654	658	1283	{kind=date, rule1=TempYear2, rule2=Ye

Annotations table





# Editing existing annotations

- **Select an annotation type from the Annotation Sets view and hover over a highlighted annotation in the text**
- A popup window displays more information about it: this is the **annotation editor**
- **Click the drawing pin symbol at the top of the editor.** This will “pin” the window open (you can still move the window around on your screen if you wish)
- **Edit the annotation: you can change the annotation type, feature names and values, the span of the annotation (clicking left and right arrows at the top of the box) or delete the annotation or its features (red Xs)**
- **Close the annotation editor by clicking the X in the top right corner, then view your edited annotation in the Annotation List**
- **Right-click a row in the annotation list → Edit... to edit that annotation**



# Annotation editor

Annotation type

The screenshot shows the GATE Developer 7.1-ANNIE interface. The main window displays a text document with several annotations. A pop-up window titled 'Location' is open, showing the configuration for this annotation type. The configuration includes a dropdown for 'locType' set to 'province', and two rules: 'rule1' set to 'Location1' and 'rule2' set to 'LocFinal'. Below the configuration is a table of annotations.

Type	Set	Start	End	Text	Feature
Date		12			
Percentvi		20			
Date		20			
Location		2108	2115	1309	{locType=province, rule1=Location1, rule2=LocFinal}
Location		2120	2125	1310	{locType=province, rule1=Location1, rule2=LocFinal}

On the right side of the interface, there is a list of annotation types with checkboxes: Date, FirstPerson, Identifier, Location, Lookup, Money, Organization, Percent, Percentvi, Sentence, SpaceToken, Split, Token, Unknown, and Original markups. The 'Location' type is checked.

feature name

value

annotation editor



# Creating new annotations

- To create a new annotation, select the portion of text you want to annotate and hover over it with the mouse.
- The annotation editor will appear: this will automatically create a new annotation.
- It will create an annotation of the same type as your last annotation: if this is your first annotation it will default to “\_New\_”. You can change this by simply editing the text.
- You can edit this annotation as before.
- You can delete the annotation by clicking on the red cross/green crayon icon
- The new annotations will appear in the currently selected annotation set. To change this, simply select a different set before creating the annotation.
- To create a new annotation set, enter a name in the text field at the bottom of the annotation sets view and click “New”.
- **Try creating some new annotations in your text.**



# Creating a Corpus

- A corpus is a collection of documents.
- For most GATE applications, it is easier to work with a corpus rather than an individual document, even if that corpus only contains one document.
- **Right click Language Resources → New → GATE Corpus**
- OR**
- **File menu → New Language Resource → GATE Corpus**
- As with the documents, you can name your corpus or use the default GATE name.



# Adding documents to a corpus

1. **With the init parameter: click the edit button  and add documents that are already loaded in GATE to the corpus. Click OK when done.**

**OR**

2. **Create an empty corpus**

**Open the corpus and use the + button to add documents, or drag them from the Resources pane**

**or populate it from a file directory (next slide)**

- **Double click on the corpus name to view the corpus.**
- **Double click the document listed there to view it.**




# Populating a Corpus (1)

- Usually, a corpus will consist of more than one document. Sometimes there could be hundreds of documents in a corpus.
- Using the populate function means you don't have to preload the documents in GATE first, and allows you to load all the documents into the corpus in one go
- To do this, let's first tidy up a bit
- It's best to keep GATE GUI clutter-free by removing any unwanted resources and documents, or it can get a bit confusing
- **Close all open documents and corpora**



# Populating a Corpus (2)

- **Create a new empty corpus, so don't add any documents to it yet**
- **Right click on the corpus name in the Resources pane and select Populate**
- **Use the file browser icon to select the name of the directory with your documents (annie-hands-on/news-texts)**
- The Extensions parameter lets you select only documents of a certain type. It is a “list” parameter where you can edit the list:
  - **Press the edit button on the right** 
  - **Type “xml” in the box (without the quotes)**
  - **Press “Add” and then “OK”**



# Populating a Corpus (3)

- “Encoding” lets you choose the right character encoding for the files. If you leave this empty, your system default is used.
  - **Enter “UTF-8” here**
- “Recurse directories” will also load documents in any subdirectories
  - **Deselect the “Recurse directories” box**
- All the documents will be loaded in one go
  - **View the contents of the corpus as before.**





# More about corpora

---

- You can use the up and down arrows to rearrange documents in a corpus
- So far the corpus and the documents are all just stored in memory. We will see later how to save documents and corpora.



# Removing documents

- To remove documents from a corpus, use the X button in the corpus editor
- Note that this does not remove the document from GATE, just from the corpus
- A document can belong to several corpora
- If you do remove the document from GATE, it will also be removed from the corpus
- But if you remove the corpus, it doesn't remove the document!
- **Experiment with adding and removing documents**



# Quick corpus creation

- If you're just testing something on one document, there's a quick way to create a new corpus and add the document to it.
- Right click on the document loaded in GATE and select “New corpus with this document”.
- This does everything in one go.
- **Try it on any document you have loaded.**
- Note that a document can belong to more than one corpus at the same time, but it can get confusing if you do this!



# Processing Resources and Plugins

- Processing resources (PRs) are the tools that process and annotate text (text processing algorithms). Often this means creating or modifying annotations on the text.
- An “application”/”pipeline” consists of any number of PRs, run sequentially over a corpus of documents
- A plugin is a collection of PRs, and other resources bundled together. For example, everything needed for IE in Arabic is in the Lang\_Arabic plugin.
- An application can use PRs from one or more different plugins.
- In order to use PRs, you need to load the relevant plugin(s)



# Plugins

---

- **Click the jigsaw icon on the top GATE menu to open the Plugin Manager  
OR use File → Manage CREOLE Plugins**



# Plugins

Load the plugin for this session only

Load the plugin every time GATE starts

List of available plugins

Resources in the selected plugin

The screenshot shows the CREOLE Plugin Manager window with the following components:

- CREOLE Plugin Manager** window title.
- Navigation tabs: **Installed Plugins**, **Available Updates**, **Available to Install**, **Configuration**.
- CREOLE Plugin Directories** section with a **Filter:** input field.
- Plugin List Table:**

	Load Now	Load Always	Plugin Name
	<input type="checkbox"/>	<input type="checkbox"/>	Alignment /media/data1_/data/work/gate-top/externals/gate/plugins/Alignment
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	ANNIE /media/data1_/data/work/gate-top/externals/gate/plugins/ANNIE
	<input type="checkbox"/>	<input type="checkbox"/>	Annotation_Merging /media/data1_/data/work/gate-top/externals/gate/plugins/Annotation_Merging
	<input type="checkbox"/>	<input type="checkbox"/>	Copy_Annots_Between_Docs /media/data1_/data/work/gate-top/externals/gate/plugins/Copy_Annots_Between_Docs
	<input type="checkbox"/>	<input type="checkbox"/>	Coref_Tools /media/data1_/data/work/gate-top/externals/gate/plugins/Coref_Tools
	<input type="checkbox"/>	<input type="checkbox"/>	Gazetteer_LKB /media/data1_/data/work/gate-top/externals/gate/plugins/Gazetteer_LKB
	<input type="checkbox"/>	<input type="checkbox"/>	Gazetteer_Ontology_Based /media/data1_/data/work/gate-top/externals/gate/plugins/Gazetteer_Ontology_Based
	<input type="checkbox"/>	<input type="checkbox"/>	GENIA /media/data1_/data/work/gate-top/externals/gate/plugins/GENIA
	<input type="checkbox"/>	<input type="checkbox"/>	Groovy /media/data1_/data/work/gate-top/externals/gate/plugins/Groovy
	<input type="checkbox"/>	<input type="checkbox"/>	Information_Retrieval /media/data1_/data/work/gate-top/externals/gate/plugins/Information_Retrieval
	<input type="checkbox"/>	<input type="checkbox"/>	Inter_Annotator_Agreement /media/data1_/data/work/gate-top/externals/gate/plugins/Inter_Annotator_Agreement
	<input type="checkbox"/>	<input type="checkbox"/>	JAPE_Plus /media/data1_/data/work/gate-top/externals/gate/plugins/JAPE_Plus
	<input type="checkbox"/>	<input type="checkbox"/>	Keyphrase_Extraction_Algorithm /media/data1_/data/work/gate-top/externals/gate/plugins/Keyphrase_Extraction_Algorithm
	<input type="checkbox"/>	<input type="checkbox"/>	Lang_Arabic /media/data1_/data/work/gate-top/externals/gate/plugins/Lang_Arabic
	<input type="checkbox"/>	<input type="checkbox"/>	Lang_Cebuano
- Resources in Plugin** panel (right side):
  - Annotation Schema
  - GATE Unicode Tokeniser
  - ANNIE English Tokeniser
  - ANNIE Gazetteer
  - Sharable Gazetteer
  - Hash Gazetteer
  - JAPE Transducer
  - ANNIE NE Transducer
  - ANNIE Sentence Splitter
  - RegEx Sentence Splitter
  - ANNIE POS Tagger
  - ANNIE OrthoMatcher
  - ANNIE Pronominal Coreferencer
  - ANNIE Nominal Coreferencer
  - Document Reset PR
  - Jape Viewer
  - Gazetteer Editor
- Buttons:** **Help**, **Apply All**, **Close**.

Apply all the settings

Close the plugins manager



# Plugins

- **Select a plugin name to see (on the RHS) the names of the resources it contains**
- **Check the “Load Now” box to load a plugin of your choice (e.g. ANNIE)**
- **Click “Apply All” to load the selected plugin**
- **Click “Close”**
- **Right click on Processing Resources to see which new PRs are now available**



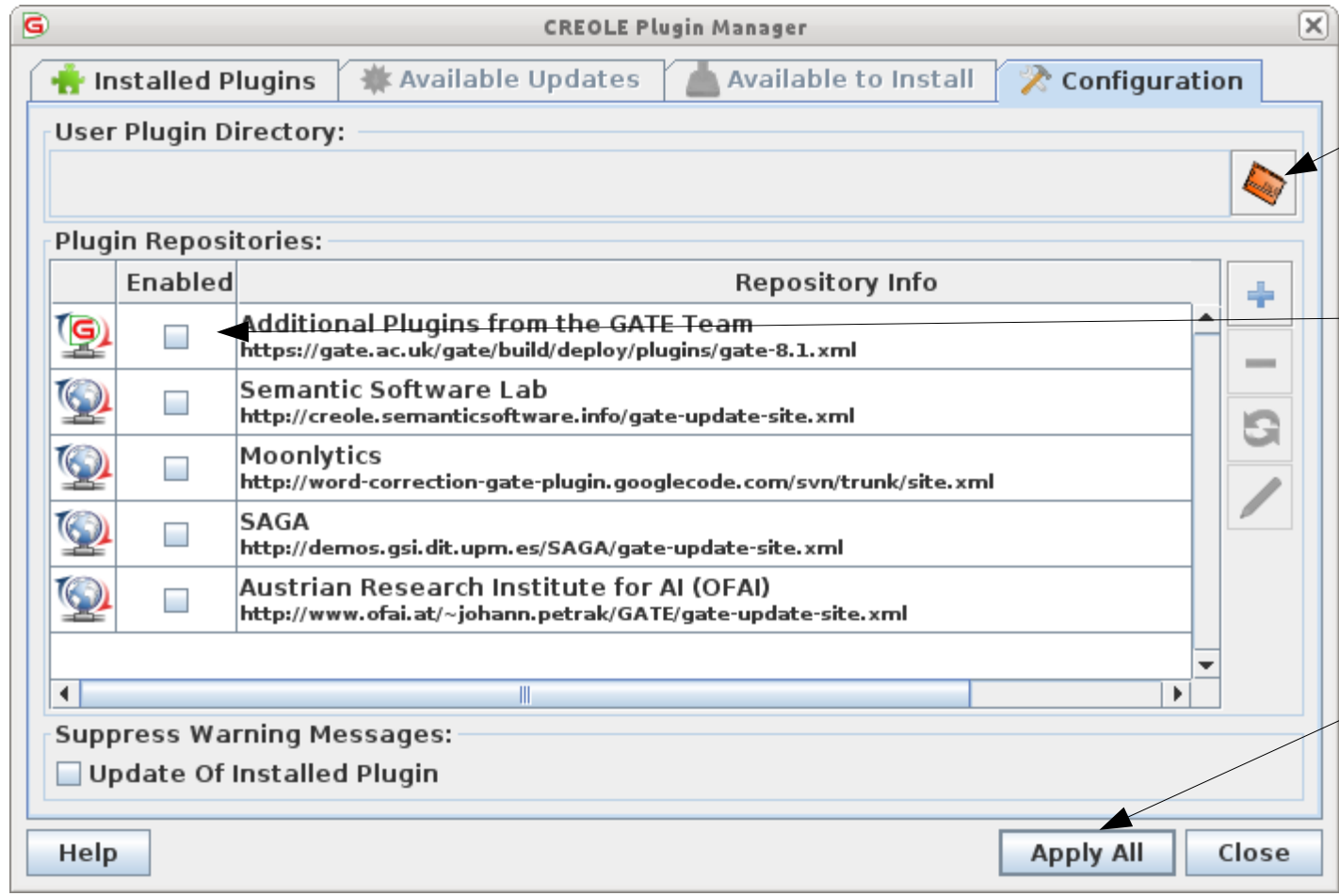
# Plugins

- GATE comes with a large number of pre-installed plugins
- Developers can provide additional plugins
- Some of the additional plugins can be installed from “Plugin Repositories” into a separate User Plugin Directory
- This is done from the “Configuration” tab and the “Available to Install” tab of the Plugin Manager
- **Show the plugin manager again**
- **Click the “Configuration” tab**





# Plugins



Choose directory  
(create new)

Choose  
Repository  
(click checkbox)

Do it!  
Get available  
plugins

Installed Plugins Available Updates Available to Install Configuration

User Plugin Directory:

Plugin Repositories:

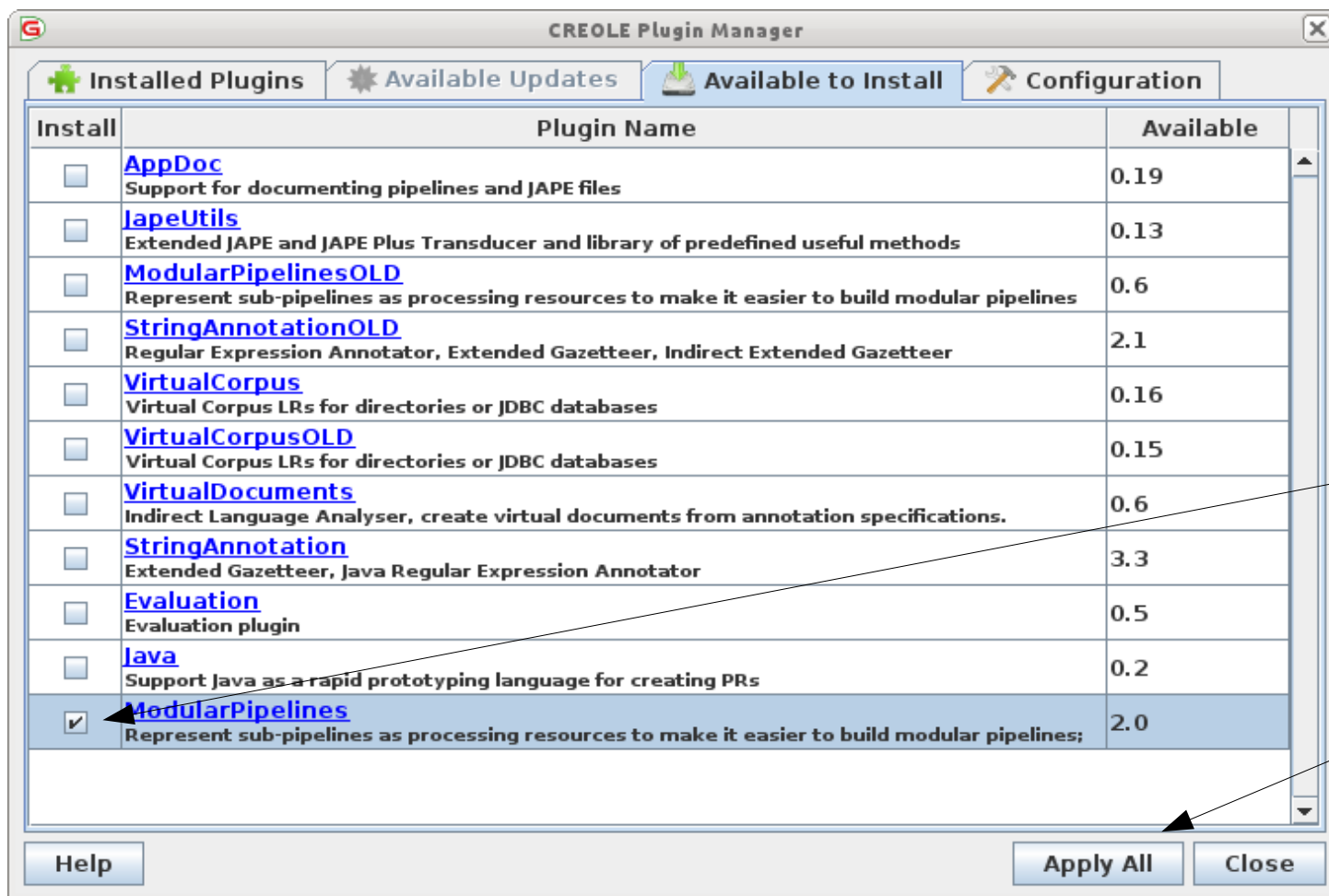
Enabled	Repository Info
<input type="checkbox"/>	Additional Plugins from the GATE Team <a href="https://gate.ac.uk/gate/build/deploy/plugins/gate-8.1.xml">https://gate.ac.uk/gate/build/deploy/plugins/gate-8.1.xml</a>
<input type="checkbox"/>	Semantic Software Lab <a href="http://creole.semanticsoftware.info/gate-update-site.xml">http://creole.semanticsoftware.info/gate-update-site.xml</a>
<input type="checkbox"/>	Moonlytics <a href="http://word-correction-gate-plugin.googlecode.com/svn/trunk/site.xml">http://word-correction-gate-plugin.googlecode.com/svn/trunk/site.xml</a>
<input type="checkbox"/>	SAGA <a href="http://demos.gsi.dit.upm.es/SAGA/gate-update-site.xml">http://demos.gsi.dit.upm.es/SAGA/gate-update-site.xml</a>
<input type="checkbox"/>	Austrian Research Institute for AI (OFAI) <a href="http://www.ofai.at/~johann.petrak/GATE/gate-update-site.xml">http://www.ofai.at/~johann.petrak/GATE/gate-update-site.xml</a>

Suppress Warning Messages:  
 Update Of Installed Plugin

Help Apply All Close



# Plugins




Choose  
Plugin

Do it!  
Install it

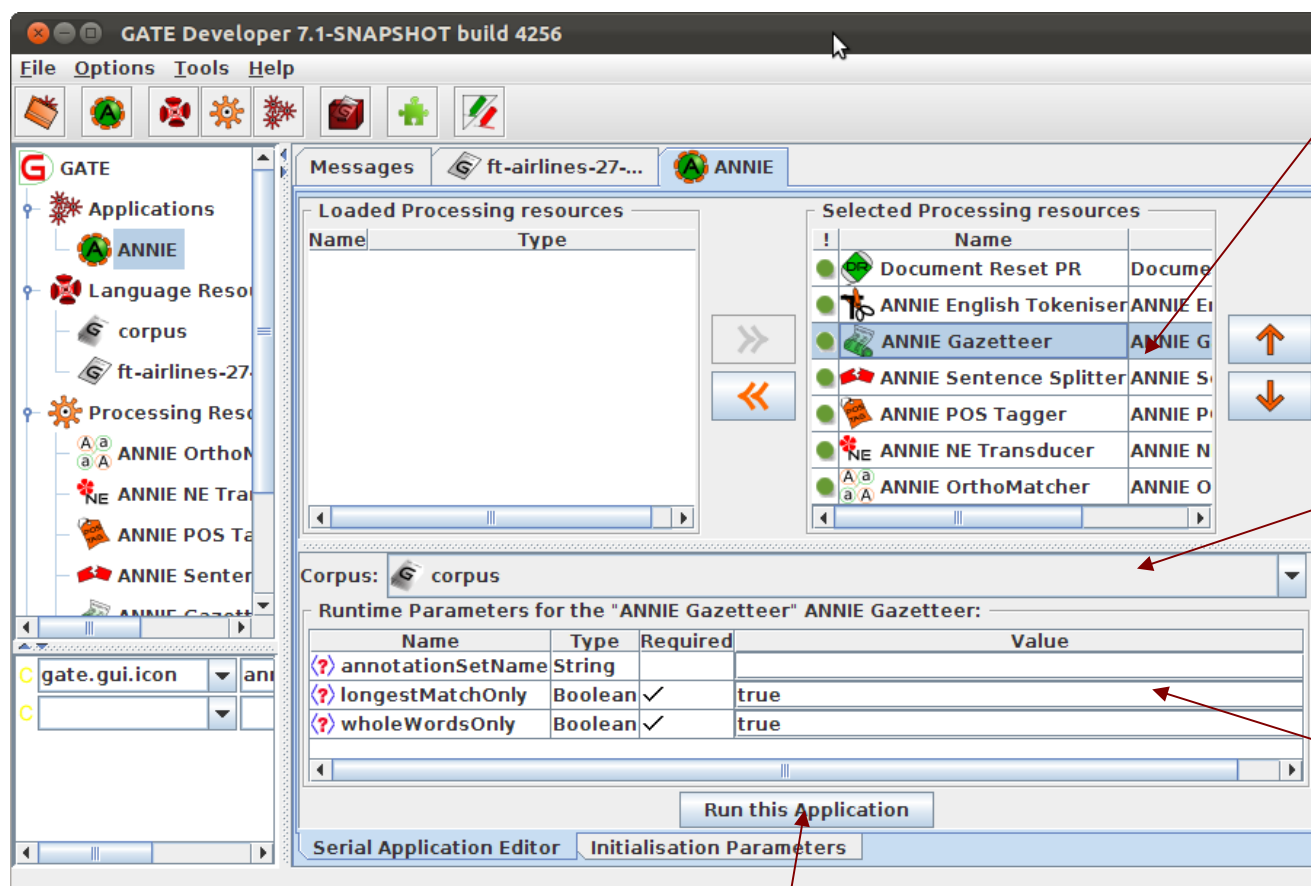


# ANNIE Application

- ANNIE is a ready-made collection of PRs that performs Information Extraction on text.
- A detailed explanation of ANNIE will be given in the second part. For now, we're just going to use it as an example of an application.
- Later, we'll show you how to create your own application
  - **Click the  icon from the top GATE menu**
  - OR Select File → Load ANNIE system**
  - **Select “with defaults”**
  - **Create a corpus with any document from the hands-on material**

# Running an application

- View the ANNIE application by double clicking on it



PRs included in application (in order of their execution)

Corpus on which the application is executed

Runtime parameters of the selected PR

Execute the application



# Running and Viewing the results

- **Choose the corpus from the drop-down list near “Corpus:”**
- **Click “Run this Application”**
- The application has finished when a message “ANNIE run in ... seconds” appears in the bottom left of your GATE window.
- **Double click on the document to view it**
- **View the annotations by selecting Annotation Sets and clicking on any Annotation types in the Default (unnamed) set**
- **If you want, you can view the annotations table too.**
- Remember that not all the results will be perfect!



# Input and Output Annotation Sets

- Some PRs use the results of previous PRs in the application. For example, the sentence splitter makes use of Token annotations produced by the tokeniser.
- The inputAS (annotation set) for the sentence splitter is the name of the annotation set where it will find the Token annotations
- The outputAS is the name of the set where it will produce the results of the sentence annotations.
- In ANNIE, the inputAS and outputAS are always the same. Later, we'll look at examples where you might want these to be different.
- Some PRs just have a parameter “annotationSetName” instead. This is because the inputAS and outputAS must be the same for that PR (usually because the PR adds information to an existing annotation rather than creating a new one)



# Changing runtime parameters

- Now we're going to change the name of the annotation set, so that all ANNIE annotations appear in a new set called ANNIEresult
- The annotation set where the results are stored is one of the runtime parameters of the PRs
  - **Double click on ANNIE to view the application and PRs.**
  - **For each PR listed, click on it and check whether it has any parameters labelled “annotationSetName”, “inputASName” or “outputASName”**
  - **Edit all of these by typing “OUT” in the box.**
  - **Double check that you haven't missed any. This is really important, otherwise your application may not work.**
  - **Now run the application again and view the results.**



# Adding new PRs (1)

- Let's add a Verb Phrase Chunker PR to ANNIE.
- First, we have to load the plugin that contains it, and then load the PR into GATE, before we can add it to the application.
- **Use the plugins manager to load the Tools plugin.**
- **Right click on Processing Resources and select “New” → “ANNIE VP Chunker”**
- **Leave all the default parameters set and click “OK”.**
- **To find out more about the VP Chunker, right click and select “Help”.**





# Adding new PRs (2)

- Now we need to add the new PR to the application.
- **Double click on ANNIE.**
- You'll see the VP chunker is in the list of loaded PRs. This means it's available in GATE, but isn't yet contained in the application.
- **Add it to the application by selecting it and using the right arrow to transfer it.**
- **Now use the up arrow to move it to the right place in the application. It should go after (below) the POS tagger but before (above) the NE transducer.**
- **Change the inputASName and outputASName parameters to ANNIEresult.**
- **Run the application and view the results on the document.**
- You should see a new annotation type “VG”.



# Saving an Application

- Applications are usually saved to files with a .gapp or .xgapp extension
- **Save your modified ANNIE application:  
Right-click the name → Save Application State**
- Saves information about the PRs, about all loaded plugins, the corpus and documents loaded into the corpus
  - Usually first remove the corpus in the application viewer
- Most locations of files referenced in the application are resolved relative to the location of the application file
- Files inside the GATE installation directory are referenced relative to that directory
- Sometimes a specific version of all plugins should get passed on with an application: “Export for GATECloud.net”
- **Close everything and re-load the saved application**



# Datstores

---

- Datstores are used to store documents and corpora on your hard disk
- Each datstore corresponds to a directory on disk
- The files stored inside the directory are a GATE specific format
- There are two kinds of datstores:
  - Serial Data Store
  - Lucene Based: not discussed now
- Running an application on a corpus that is stored in a datstore will automatically save each processed document
- With datstores, only the document needed is loaded into memory



# Create a new Serial Datastore

- **Right click “Datastores” from the Resources pane and select “Create Datastore”**
- **Select “Serial Datastore”**
- **Create a new empty directory by clicking the “Create New Folder” icon and give your new directory a name**
- **Select this directory and click “Open”**
- Now your datastore is ready to store your documents



# Save documents to the datastore

- **Right click on your corpus and select “Save to Datastore”**
- **Select the datastore that you just created**
- **Now close the corpus and document**
- **Double click on the name of the datastore in the Resources pane**
- You should see the corpus and document
- **Double click them to load them back into GATE and view them**
- They should contain the annotations you created previously
- You can remove things from the datastore by right clicking on their name in the datastore and selecting “Delete”
- You can add several corpora to the same datastore



# If you have lots of documents..

- A datastore is the best way to store them, because it uses less memory in GATE when processing
  - **Delete all corpora and documents in your datastore**
  - **Load a new corpus**  
**(Language Resources → New → GATE Corpus)**
  - **Create a new datastore and save the (empty) corpus in it**
  - **Now populate your corpus (right click on corpus → Populate)**
- You should see the documents appear in your datastore  
Your documents will be loaded into the datastore and saved automatically.
  - **Close and reopen your datastore to check they really were saved!**



# Saving documents outside GATE

- Datastores can only be used inside GATE, because they use a GATE-specific format
- If you want to use your documents outside GATE, you can save them in 2 ways:
  - Using stand-off annotations, in GATE format (GATE XML .xml)
  - using “inline” annotations (Inline XML .xml): only non-overlapping annotations can be saved
- Both formats are XML-based
- With the “Format\_FastInfoSet” plugin loaded you can save them in a compressed XML format (Fast InfoSet .finf)



# Saving as XML

- **Load any document from the hands-on material into GATE, then right click on it in the Resources pane**
- **Select “Save as → GATE XML” and select a filename.**
- In this GATE-specific format, all annotations are kept separate from the text.
- **If you’re curious, load the document into your favourite text editor and have a look at it!**





# Save preserving format

- This option will save the document with all the original annotations from HTML or XML documents, and any new annotations that you currently have selected in the document editor
- This can be useful for saving only selected annotation types
- Annotations are saved using standard XML tags, with the annotation type as the tag name
- ***Partially overlapping annotations can not be saved!***
- **Right click and use “Save as → Inline XML”**
- If the Advanced Option in GATE “Include annotation features for save preserving format” has been checked, then selected features will be saved as well as annotations, in this mode.



# Setting up GATE options

- You can set up different options in GATE using the Options menu.
- **Click Options** → **Configuration** → **Appearance** to change the look and feel of GATE, such as menu and text fonts
- Clicking the Advanced tab enables you to adjust settings such as saving your options, and saving the session so that when you reopen GATE, it will remember and reload the applications you had open at the end of your previous session



# Advanced options

---

- Save options on exit: save your options in a file when GATE exits (`~/ .gate.xml` on Linux)
- Save session on exit: save loaded applications and restore next time (`~/ .gate.session`)
  - Useful for reloading the pipelines and plugins one was working on last.
  - Avoid when running several copies of GATE at the same time (but possible to use separate configuration files per project, see manual)
- Link resources ... : when a resource is selected in the viewer it gets automatically selected in the resource pane or vice-versa.



# Summary

---

- This tutorial has given you a guided tour of the GATE GUI
- Looked at
  - language resources
  - datastores
  - applications
  - processing resources
- There are lots of other tools and options you can play with: see the User Guide for more info
- In the next session we will look at the topic of Information Extraction and further examine ANNIE, GATE's default IE system



# Extra exercises

---

If you have some spare time, you can try some more exercises:

- **Load an HTML or XML document with the markupAware parameter set to false and see the difference**
- **Investigate the AnnotationStack**
- **Play with Advanced Options**
  - **Advanced Features**
  - **Session persistence**
  - **Linking resources**
- **Run an application over documents in a datastore**