

Module 1: Information Extraction

Part 1: Introduction to GATE

David Jones
Mehmet Bakir



About this tutorial

- This tutorial will be a hands on session with some **explanation as you go**.
- As topics are introduced, there'll be time for you to try playing with different parts of the GUI.
- Things for you to try yourself are in **red**.
- There'll be extra time at the end to practise again, or go on to some further exercises. 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, you'll learn more about the underlying functionality. So please reserve your burning questions about this for a little bit longer!

Why GATE?

- GATE is the most widely used open source toolkit for NLP in the world
- We're using it because it's a great way to showcase all the core NLP components that are used for text analysis tasks
- You can play with all the tools in GATE and try out things for yourself to see how it works
- And also because we're experts
 - Developed at the University of Sheffield since 2000 (in its current form)
 - The people who developed the NLP tools in GATE since 2000 are the ones presenting this course 😊
- And by the way, just because it's old doesn't mean it's out of date. GATE is in constant development with new technologies being constantly added.

What is 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)
- Scalable from laptop to massive processing on the cloud (including real-time stream processing)

GATE 8.6

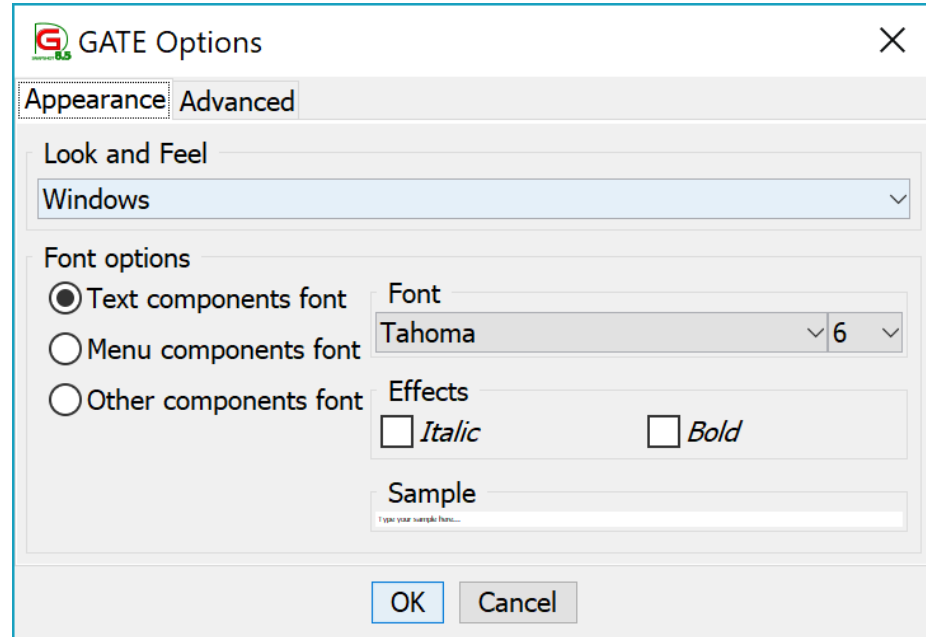
- In this course, we will use GATE 8.6. If you have an older version, please upgrade to the newer one, or many things may not work.
- This version was just released last week – with any luck we have fixed all the bugs 😊
- In the latest version, we made a few changes, mainly for routine maintenance.
- We provide a way to upgrade existing applications though, and the new way is much better overall.
- **Start GATE on your computer now (if you haven't already)**

Time to get your hands dirty!



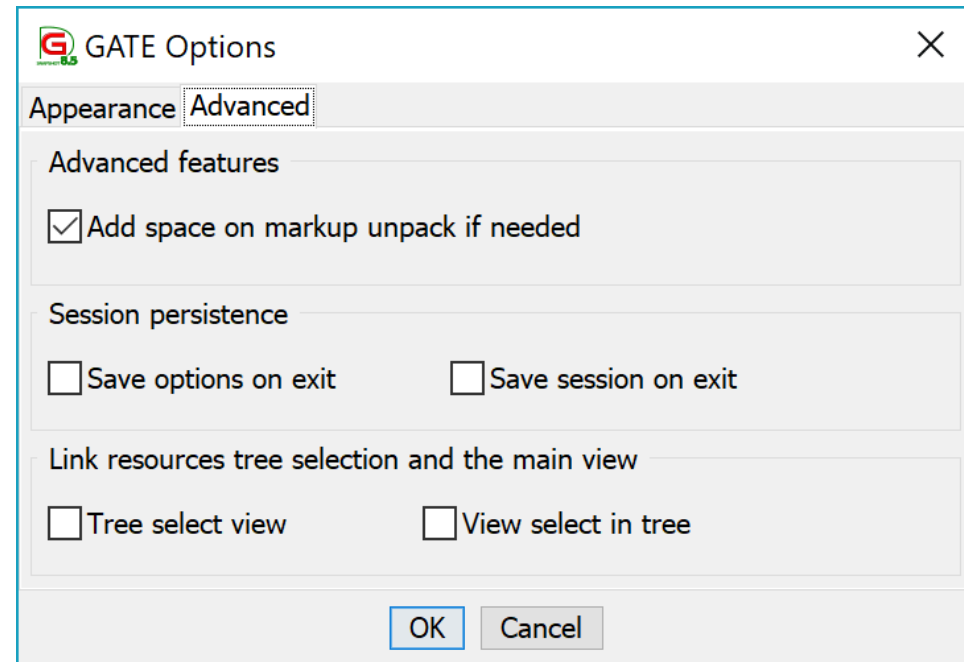
Setting up GATE options

- 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
- We recommend the Metal Look and Feel (depending on your OS, some features may not work with others)



Setting up GATE options (2)

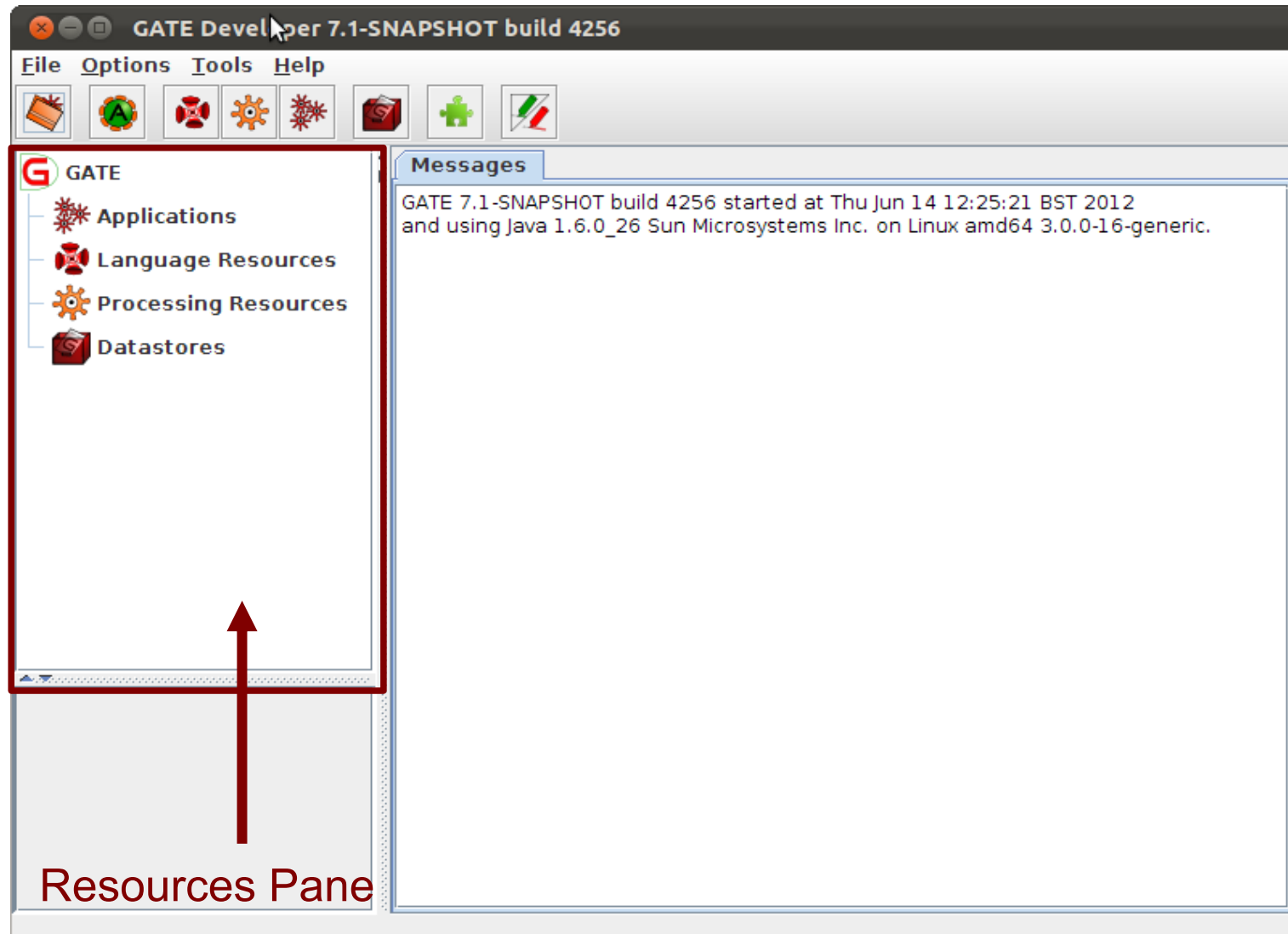
- Clicking the Advanced tab enables you to adjust settings such as saving your options, and saving the session so that when you reopen GATE, it will remember and reload the applications you had open at the end of your previous session.
- You can try this out later.



1. Finding your way around the GUI

- How to navigate the GATE GUI
- How to set up the different options
- Introduction to resources and parameters

Resources Pane



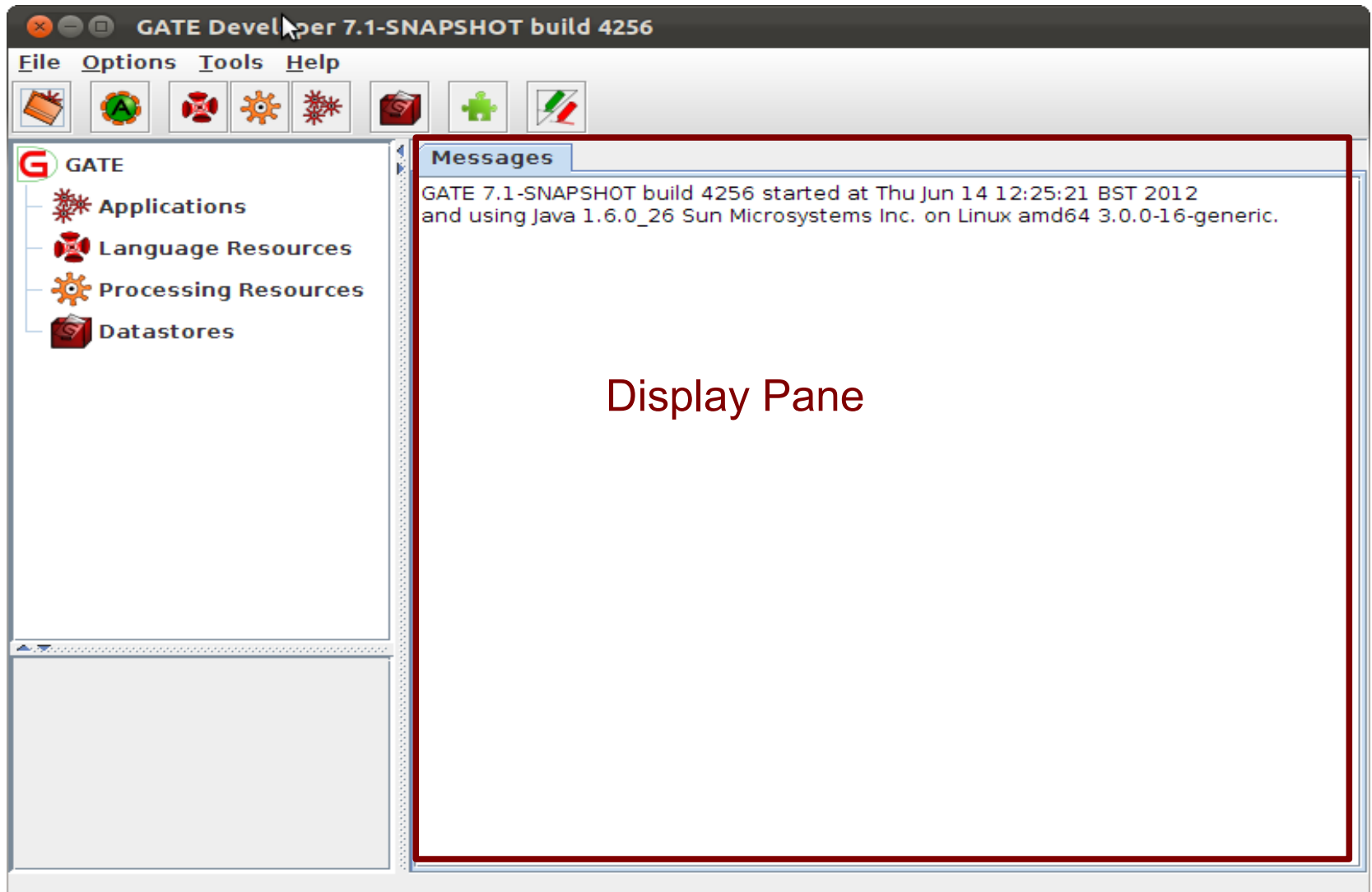
Resources Pane

- **Language resources** (LRs) are documents or document collections
 - a collection of documents is known as a **corpus**
- **Processing resources** (PRs) are annotation tools that operate on text within the documents
- **Data stores** are specialised files where documents are kept for future use
- **Applications** are groups of processes that run on one or more documents

Simple operations on resources

- In general, right clicking on the name of a resource in the resource pane gives access to a menu of actions
- Double clicking on an instance of a resource enables you to view the resource
- Selecting a resource instance and pressing Delete will generally close it
- You can also right click and then select “Close”

Display Pane



Displaying Elements

- When you first open GATE, the Display page will typically just display any messages from the system
- It displays whatever elements you are currently working with, e.g. an application, a document or a processing resource
- Double clicking on an instance of any resource will generally display it
- Along the top of the pane may be various tabs which allow you to toggle the views of any open resources
- Clicking on a tab displays that view
e.g. “Messages” tab shows messages

Parameters

- Applications, LRs, and PRs all have various parameters which can be set either at load time (initialisation) or at run time.
- Parameters enable different settings to be used, e.g. case sensitivity
- **Initialisation Parameters** (set at load time) cannot be changed without reloading (these may be called “init parameters” for short)
- **Run time Parameters** can be changed between each application run
- Later you'll be able to experiment with setting parameters on resources and applications

2. Loading and Viewing Documents

- Loading a document and setting its parameters
- Navigating through documents and viewing their annotations

Loading a document

- When GATE loads a document, it converts it into a special format for processing
- GATE can process documents in all kinds of formats: e.g. plain text, HTML, XML, PDF, Word.
- Documents have a markupAware parameter which is set to true by default: this ensures GATE will process any existing annotations such as HTML tags and present them as annotations rather than leaving them in the text.
- Documents can be exported in various formats or saved in a datastore for future processing within GATE

Loading documents

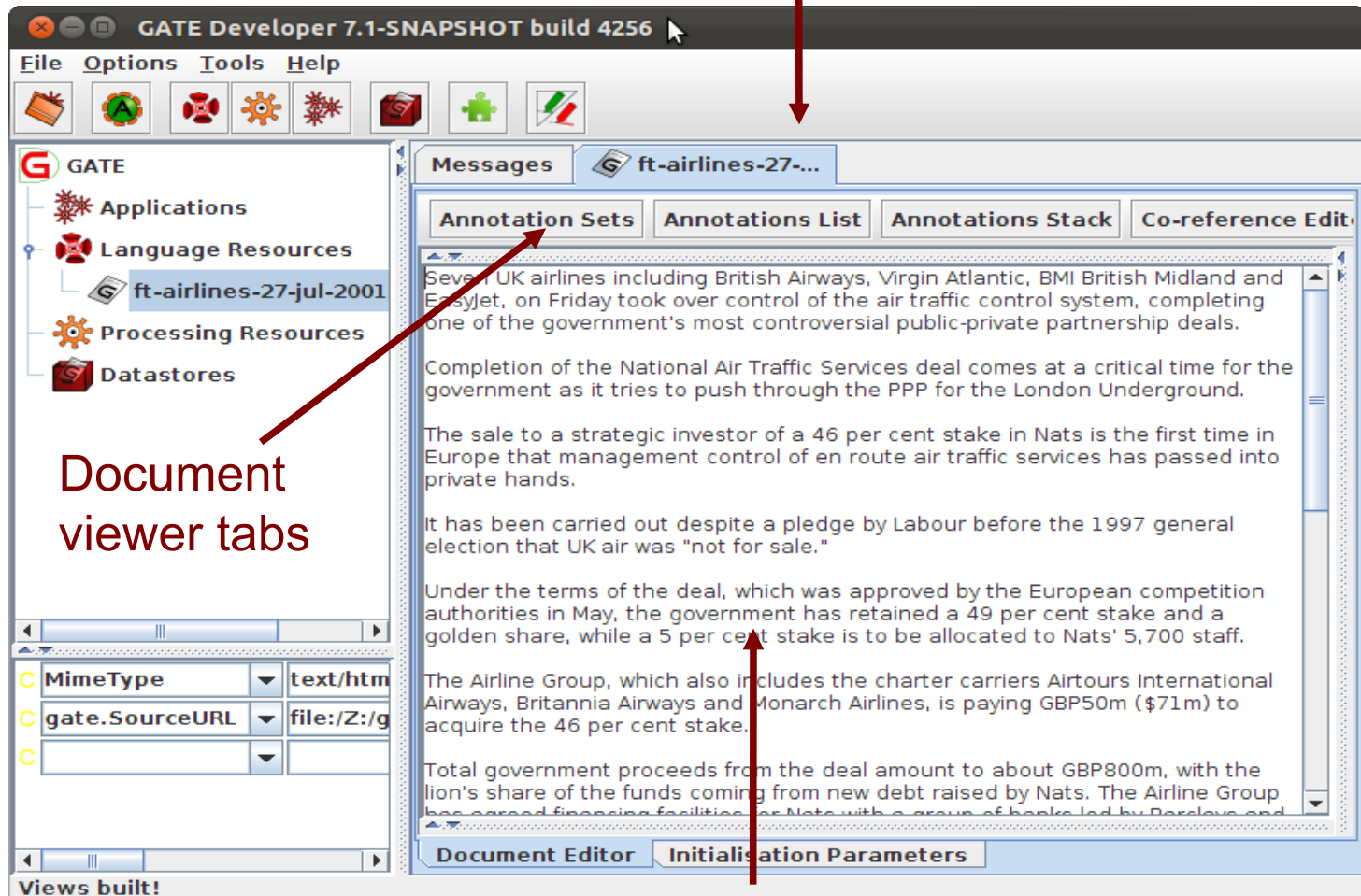
- To load a document, you can right click on Language Resources and select “**New → GATE Document**”
- You can also go via the **File menu → New Language Resource → GATE Document**
- The **sourceURL** parameter enables you to specify the document to be loaded. You can type the **filename** or **URL**, or click the **file browser icon** to navigate to the correct document.
- Try loading a file from your hands on materials *and* one from the Web – you must include **http://** when specifying a URL
- You can also just type a string of text into the box. In this case, you need to select **stringContent** rather than **sourceUrl**, using the arrow, before typing the text.

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 other kinds of resources such as PRs

Document viewer

Highlighted tab is the resource currently being viewed



Opening and closing documents

- To view a document, double click on the document name in the Resources pane
- To close a document, right click on the document name and select **“Close”**
- To hide a document, while leaving it loaded, right click on the document tab and select **“Hide”**
- The Document viewer buttons at the top of the Display pane let you select different views
- To view the annotations, you first need click **“Annotation Sets”**, and then select the relevant set and annotation(s) on the right
- To see a list of annotations at the bottom, click on **“Annotations List”**
- Load the **“ft-airlines-27-jul-2001.xml”** file from your hands-on folder

3. All about Annotations

- Introduction to annotations, annotation types and annotation sets
- Creating and viewing annotations

Annotations

- The annotations associated with each document are a structure central to GATE
- Each annotation consists of
 - start and end offsets
 - optionally a set of features associated with it
 - each feature has a name and a value

Annotation Sets

- Annotations are **grouped into sets**, e.g. Default, Original Markups
- Each set can contain a number of **typed annotation**, e.g. Person, Location etc.
- You can create and organise your annotation sets as you wish.
- It's useful to **keep different sets for different tasks** you may perform on a document, e.g. to separate the original HTML tags from your new annotations
- It's important to **understand the distinction between annotation set, annotation type, and annotation**
- This is best explained by looking at them in the GUI

Annotation Sets

The screenshot displays the GATE Developer 8.6 build 86a246c interface. The left sidebar shows the project structure under 'GATE', including 'Applications', 'Language Resources', 'Processing Resources', and 'Datastores'. The central text editor displays a news article about UK airlines. The right sidebar shows the 'Annotation Sets' list, which includes 'Key' and 'Original markups'. Red arrows point from text labels to these elements:

- 'Key' annotation set points to the 'Key' annotation type in the list.
- Annotation types points to the 'Original markups' annotation type in the list.
- Original Markups annotation set points to the 'Original markups' annotation type in the list.

The text editor displays the following text:

Seven UK airlines including British Airways, Virgin Atlantic, BMI British Midland and Easyjet, on Friday took over control of the air traffic control system, completing one of the government's most controversial public-private partnership deals.

Completion of the National Air Traffic Services deal comes at a critical time for the government as it tries to push through the PPP for the London Underground.

The sale to a strategic investor of a 46 per cent stake in Nats is the first time in Europe that management control of en route air traffic services has passed into private hands.

It has been carried out despite a pledge by Labour before the 1997 general election that UK air was "not for sale."

Under the terms of the deal, which was approved by the European competition authorities in May, the government has retained a 49 per cent stake and a golden share, while a 5 per cent stake is to be allocated to Nats' 5,700 staff.

The Airline Group, which also includes the charter carriers Airtours International Airways, Britannia Airways and Monarch Airlines, is paying GBP50m (\$71m) to acquire the 46 per cent stake.

Total government proceeds from the deal amount to about GBP800m, with the lion's share of the funds coming from new debt raised by Nats. The Airline Group has agreed financing facilities for Nats with a group of banks led by Barclays and Abbey National.

Completion of the deal has come about two months behind the original schedule announced at the end of March.

It is understood that negotiations were held up by concerns expressed by the banks financing the deal about revised traffic forecasts presented by Nats after the selection of the Airline Group as the government's partner was announced at the end of March.

The bottom of the interface shows a table with columns: Type, Set, Start, End, Id, Features. Below the table, it says '0 Annotations (0 selected) Select:'. At the bottom, there are buttons for 'Document Editor', 'Initialisation Parameters', and 'Relation Viewer'.

Viewing annotations

- **Double click on your document** to view it
- **Click on the Annotation Sets** button to open a new pane on the right hand side (Annotation Sets view)
- **Default (unnamed) set** contains some examples of annotations
- **Click on the arrow** to display the annotation types belonging to that set
- You should see types such as Location, Date, Person etc.
- **Select an annotation type** to view all the annotations of that type in the document

A closer look at the annotations

- Select the **Annotations List** button from the menu above the Display pane
- For each annotation type selected in the Annotation sets view, all annotations corresponding to that type will be shown in the table
- Table shows annotation type, offsets, annotation set, features and values
- Select a row in the table to highlight the annotation in the text
- Click on a column heading to sort according to the header
- We will look at **AnnotationStack** and **Coreference Editor** views later

Annotations

Date annotation

The screenshot shows the GATE Developer 7.1-SNAPSHOT build 4256 interface. The main window displays a text document with several paragraphs. Annotations are visible in the text, including "last year", "5 per cent", "2000", "England", "Wales", "next January", "next 10 years", and "UK". A red arrow points from the "Date" annotation in the table below to the "next 10 years" annotation in the text.

Annotations table:

Type	Set	Start	End	Id	Feat
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	1242	{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)
- **Try editing 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

Annotation editor

The screenshot shows the GATE Developer 7.1-ANNIE interface. The main window displays a text document with several annotations. A red arrow points to the 'Annotation type' label, which is positioned above the 'Annotations List' tab. Another red arrow points to the 'feature' label, which is positioned below the 'Annotations List' tab. A third red arrow points to the 'value' label, which is positioned below the 'Annotations List' tab. A fourth red arrow points to the 'Annotation editor' label, which is positioned below the 'Annotations List' tab. The 'Annotations List' tab shows a table of annotations with columns for Type, Set, Start, End, and Value. The 'Annotation editor' tab shows a form for editing annotations, including fields for locType, rule1, rule2, and a 'New' button.

GATE Developer 7.1-ANNIE build 4256

File Options Tools Help

Annotations List

Annotation type

feature

value

Annotation editor

Type	Set	Start	End	Value
Date	19			
Percentvi	20			
Date	20			
Location	2108	2115	1309	{locType=province, rule1=Location1, rule2=...
Location	2120	2125	1310	{locType=province, rule1=Location1, rule2=...


4. Documents and Corpora

- Creating and populating a corpus of documents in different ways

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.

Ways to add documents to a corpus

1. **Click the edit button**  and add the documents that are already loaded in GATE to the corpus
 - **Click OK**
2. **OR**
 - **Create an empty corpus**
 - **Double click on the corpus name** to open the corpus
 - **use the + button** to add documents, or **drag them from the Resources pane**
 - **Double click the document listed there** to view it.
3. **or** populate it from a file directory (next slide)

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** as before, so don't add any documents to it yet
- **Right click on the corpus name** in the Resources pane and **select Populate**
- **Select the name of the directory** with your documents
- The **Extensions parameter** lets you select only documents of a certain type.
 - **Press the edit button**
 - **Type “xml”** in the box (without the quotes), **press “Add”** and then **“OK”**
- **“Encoding”** lets you choose the right encoding for the documents. The wrong encoding can cause characters to be incorrectly displayed: **Enter “UTF-8”**
- **“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
- Click on the tab at the bottom to **view the initialisation parameters** of a corpus

Cheat's tip for 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 clicking on the document** loaded in GATE and **selecting “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!


5. Processing Resources and Plugins

- Loading processing resources and managing plugins

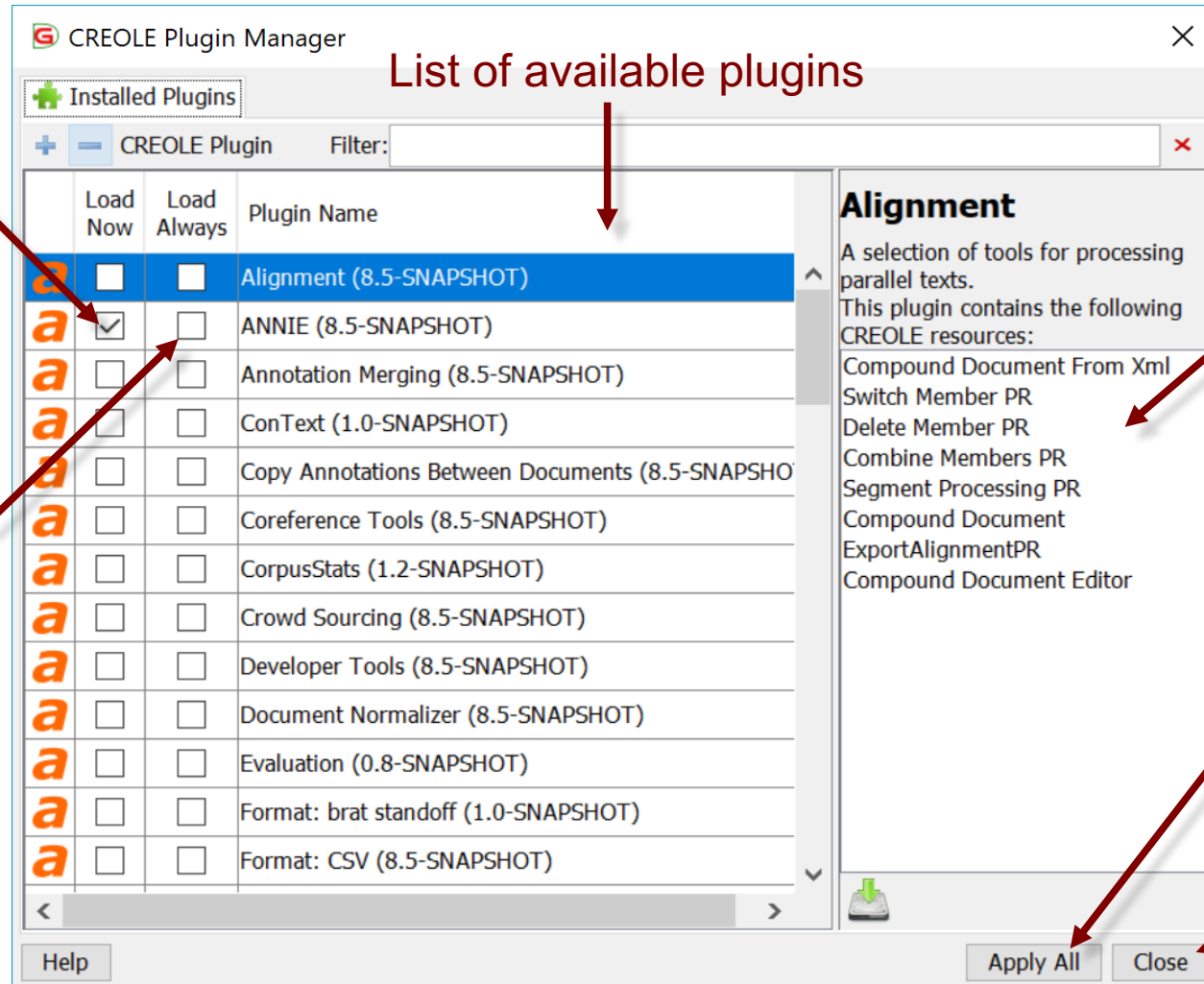
Processing Resources and Plugins

- **Processing resources (PRs)** are the tools that enable annotation of text. They implement and execute text processing algorithms. Typically this means creating or modifying annotations on the text.
- **An application consists of any number of PRs**, run sequentially over a corpus of documents.
- **A plugin is a collection of one or more PRs**, bundled together. For example, all the PRs needed for IE in Arabic are found in the Lang_Arabic plugin.
 - A plugin may also contain language or visual resources, but you don't need to worry about that now!
- An application can contain PRs from one or more different plugins.
- **In order to access new PRs, you need to load the relevant plugin.**

Plugins

- Warning: plugins have changed since the last version of GATE!
- Click the  icon on the top GATE menu to open the Plugin Manager [or go via File → Manage CREOLE Plugins]
- You should see a popup box appear with a list of plugins (this may take a few seconds the first time)
- Click on a plugin name to see the information about it

Plugins



Load the plugin for this session only

Load the plugin everytime GATE starts

List of available plugins

Resources in the selected plugin

Apply all the settings

Close the plugins manager

Apply All


Close

Help

Plugins

- Select a plugin to see (on the RHS) the names of the resources it contains
- Check the relevant “**Load Now**” box to load a plugin of your choice
- Click “**Apply All**” to load the selected plugin(s) (This may take a few seconds the first time you do it)
- Click “**Close**”
- Right click on **Processing Resources** to see which new PRs are now available (this should match the list you saw in step 1!)


Downloading the resources from a plugin

- If you want to edit some of the resources in a plugin (for example, a gazetteer list from ANNIE) you need to first extract them
- In the plugin manager, **select ANNIE, click the download button**  **and save it to a local folder**
- Remember where this local folder is – you will need it later!

6. Applications

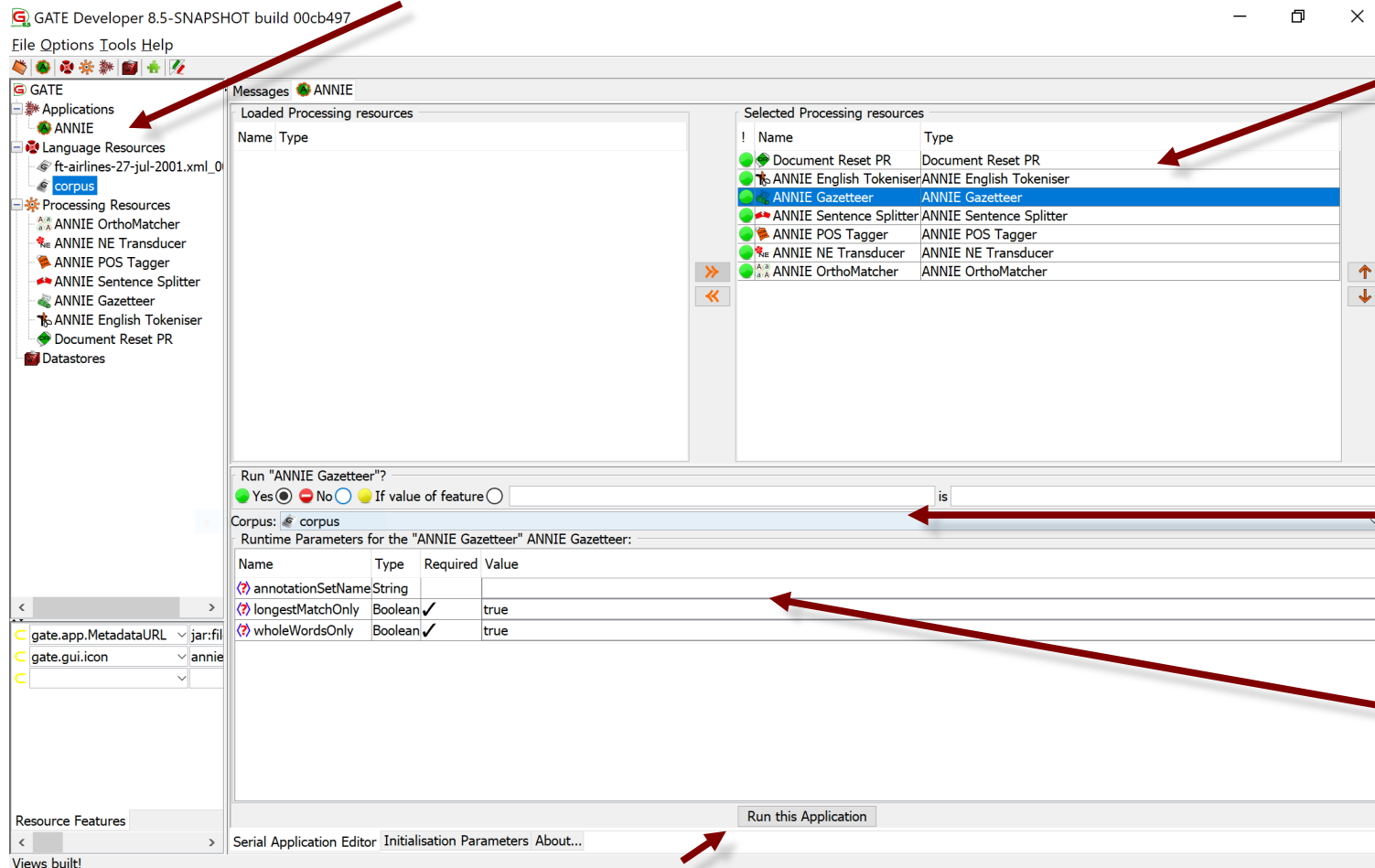
- Loading and running ANNIE and pre-existing applications
- Creating a new application

Here's one we made earlier: ANNIE

- ANNIE is a ready made collection of PRs that **performs Information Extraction on unstructured 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 make your own application from scratch.
- **Click the  icon from the top GATE menu OR Select File → Ready Made Applications → ANNIE → ANNIE**
- **Load any document from the hands-on material and add it to a corpus**

Running an application

1. View the ANNIE application by double clicking on it



2. PRs selected in application (in order of their execution, don't change for now.)

3. Select the corpus on which the application is executed

4. Runtime parameters of the selected PR

5. Execute the application

Viewing the results

- After the application finished.
- **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! Later in the course, you'll learn more about the causes of these errors.

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 “ANNIEresult” 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 (8.6) plugin.**
- **Right click on Processing Resources and select “New” → “ANNIE VP Chunker”**
- **Leave all the default parameters set and click “OK”**

Adding new PRs (2)

- Now we need to **add the new PR to the application**.
- **Double click on ANNIE.**
- You'll see the **Annie 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”**.

7. Saving documents

- Using datastores
- Saving documents for use outside GATE

Types of datastores

There are 2 types of datastore:

- **Serial datastores** store data directly in a directory.
- **Lucene datastores** provide a searchable repository with Lucene-based indexing.

For now, we'll look at serial datastores. Lucene is covered on Friday/Applications track.

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
- Note: if this icon does not appear, try selecting the Metal Look & Feel (especially Mac users)
- **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 on 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
- Note: in general, it's best to save the empty corpus to the datastore and **then** populate it, to avoid keeping a lot of documents in memory

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:
 - **Gate XML is a standoff markup**, a special GATE representation
 - **Inline XML has inline annotations** (preserving the original format)

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 format, all annotations are appended to the end of the document and the location for each annotation is marked by a tag in the body of the document
- Each annotation has a unique ID
- If you're curious, load the document into your favourite text editor and have a look at it!

Summary

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

Optional Material

If you have lots of documents in a corpus...

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** to the datastore
- 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!

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.
- 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.

Save as Inline XML

- 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 a subset of the 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 on a document and select “**Inline XML**”
- Enter the annotationSetName
- Select annotationTypes
- Enter a name for the rootElement, e.g. doc
- Make sure the target path in Save To is correct (this can be an issue in windows machines)
- Click OK to save it.

Removing documents from corpora

- 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
 - The document is available to be added to other corpora. Indeed a document can belong to several corpora
- If you do remove the document from GATE, it will also remove it from the corpus
 - But if you remove the corpus, it doesn't remove the document!
- Try experimenting with adding and removing documents

Further 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
- Run an application over documents in a datastore