Module 1: Introduction to GATE Developer

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.

- Start GATE on your computer now (if you haven't already)

- 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!
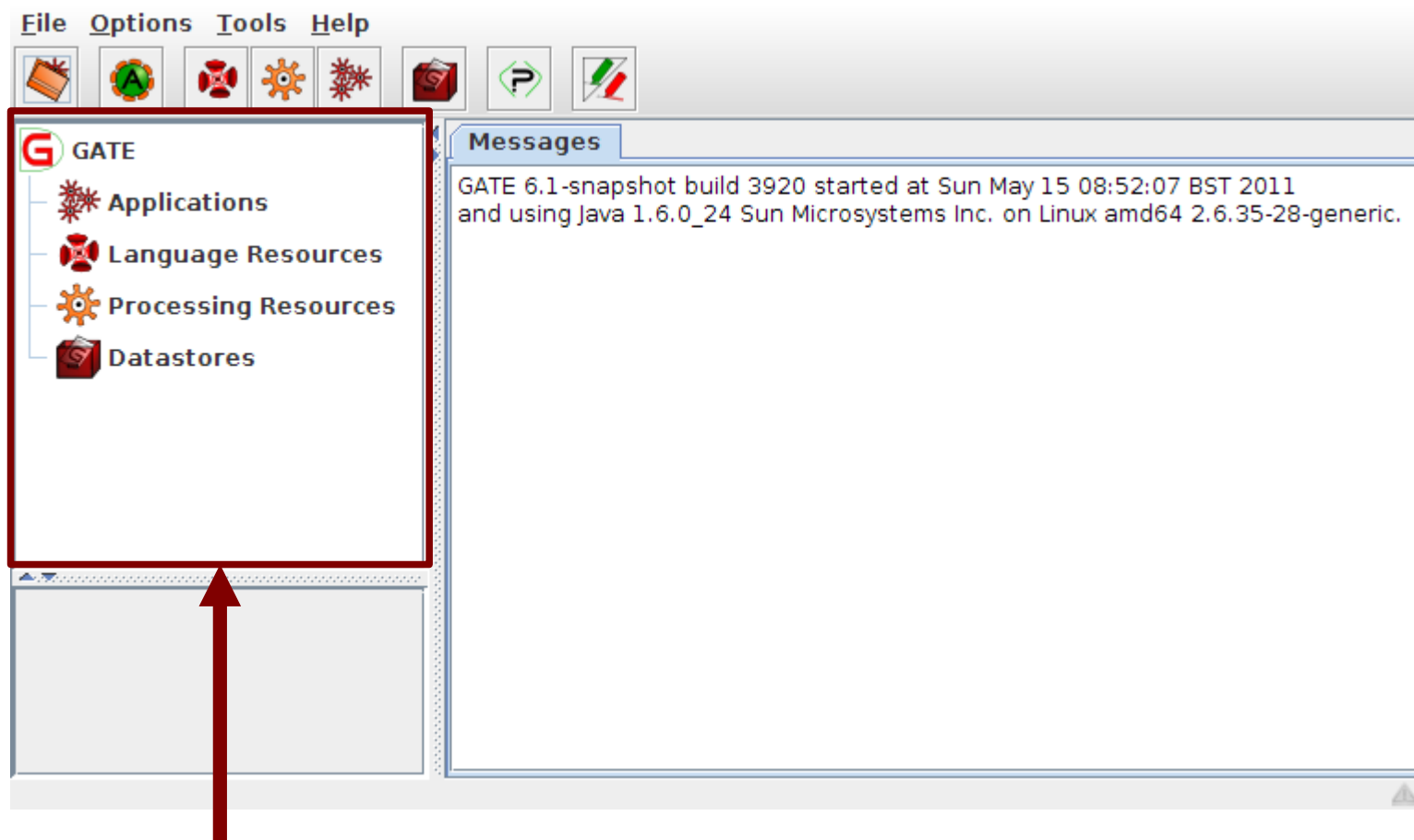
Time to get your hands dirty!

1. Finding your way around the GATE GUI

- How to navigate the GATE GUI

- How to set up the different options

- Introduction to resources and parameters

Resources Pane

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
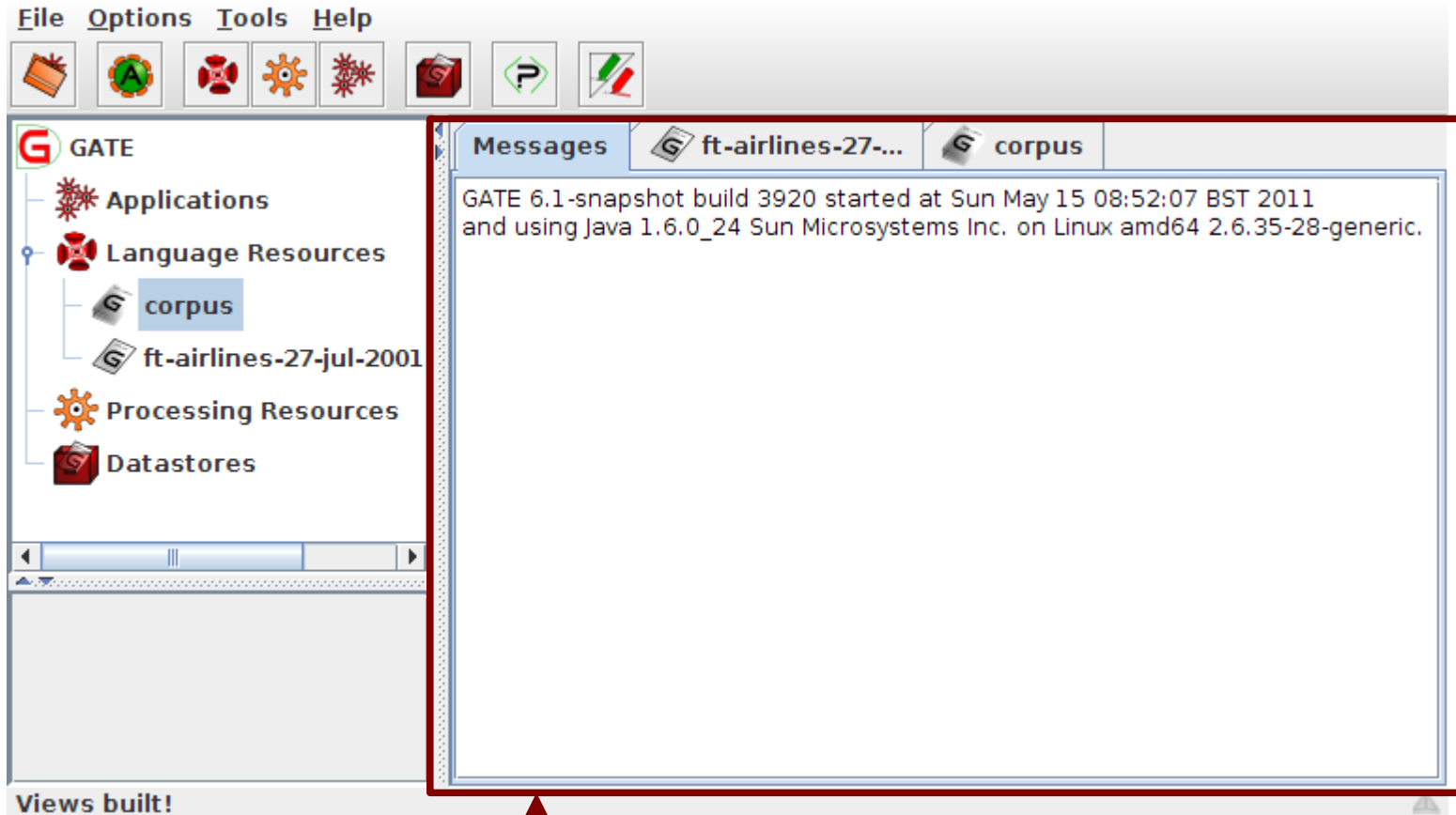
- You can also right click and then select "Close"

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

Display Pane

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

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

  ➢ Try a few different options.

- 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.

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: plain text, HTML, XML, PDF, Word etc.

- 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 the http:// part 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.
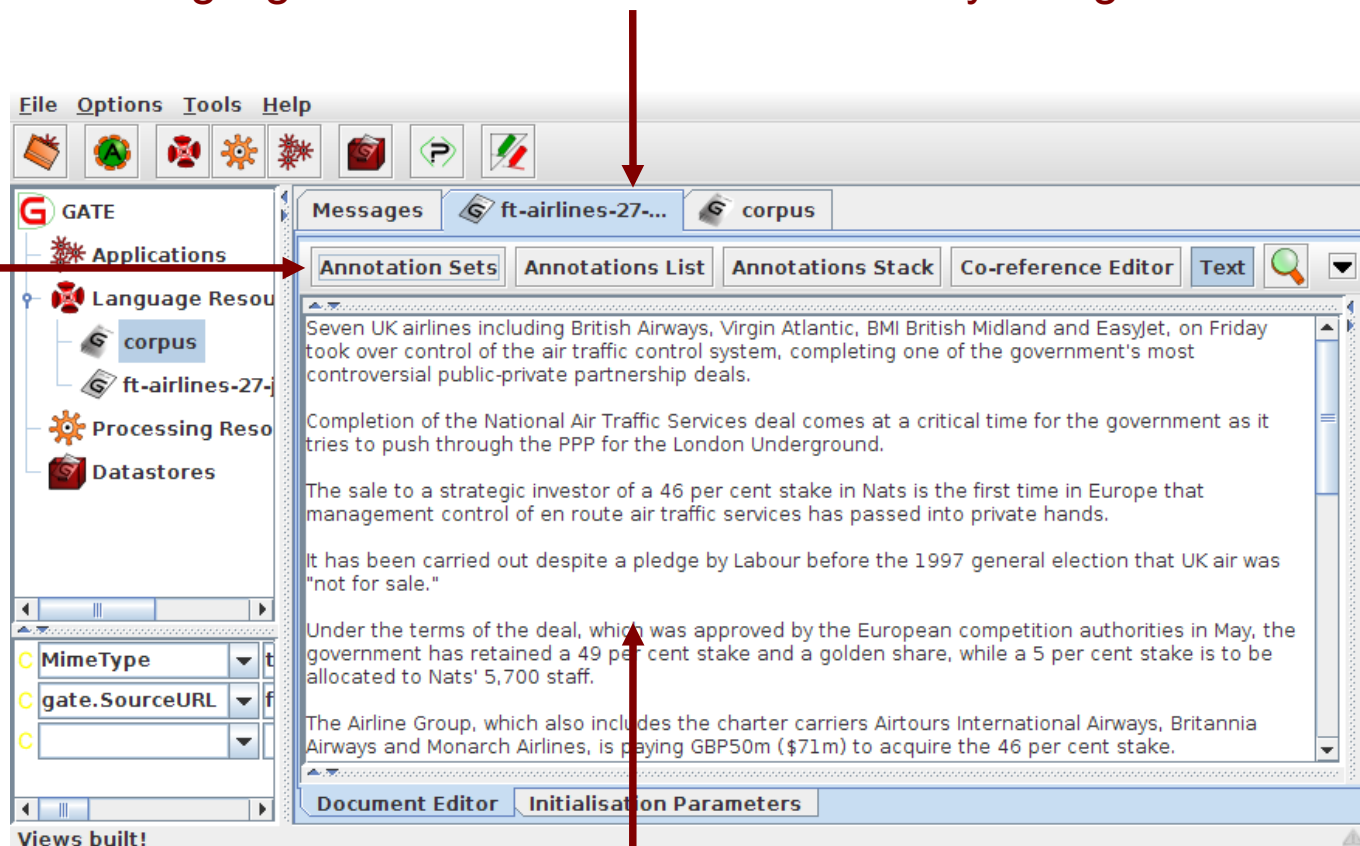
- Try loading a document via the stringContent method

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

Document viewer buttons

Document

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 one of the HTML files in 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
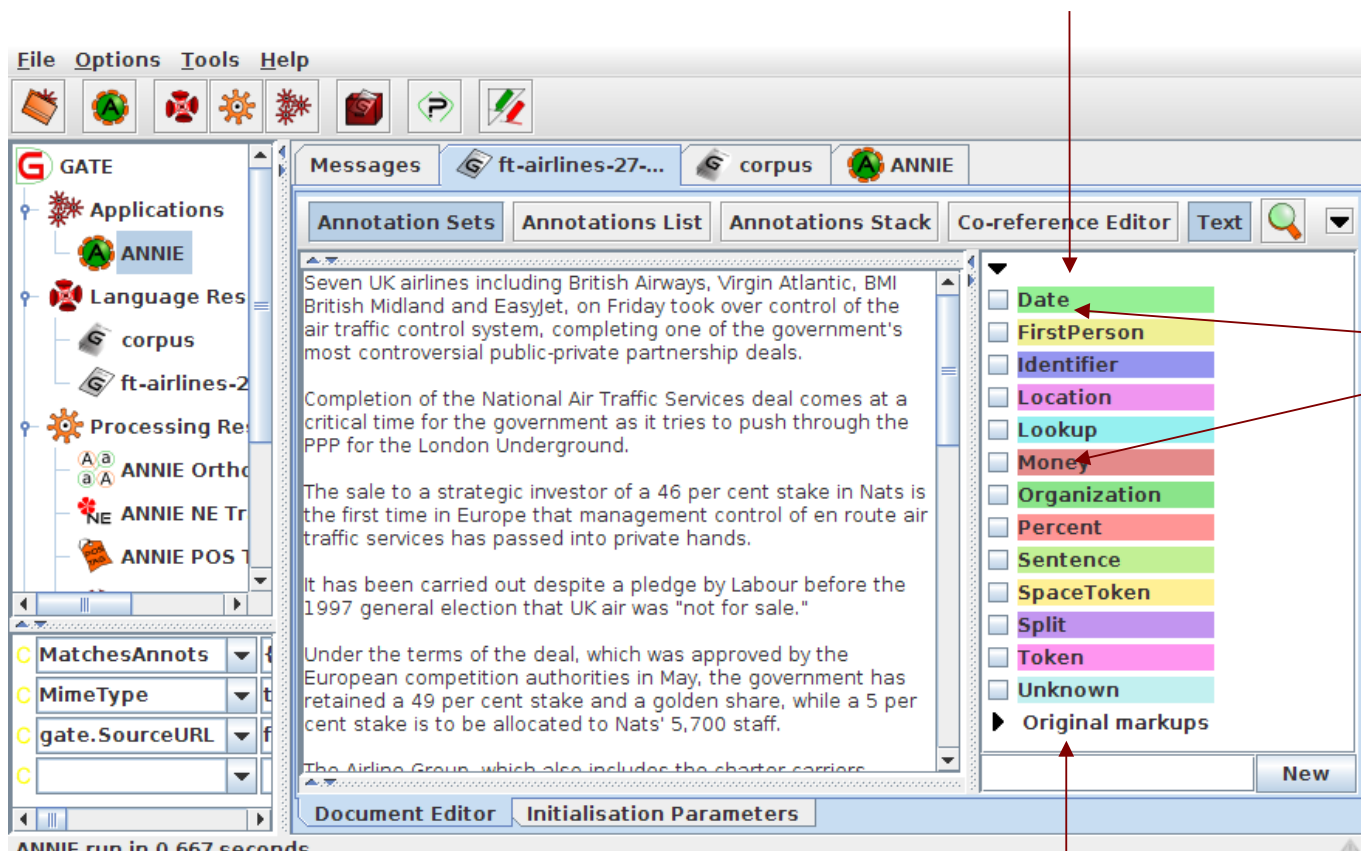
  - each feature has a name and a value

Annotation Sets

- Annotations are grouped into sets, e.g. Default, Original Markups

- Each set can contain a number of annotation types, 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

Default annotation set

Annotation types

Original Markups annotation set

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.
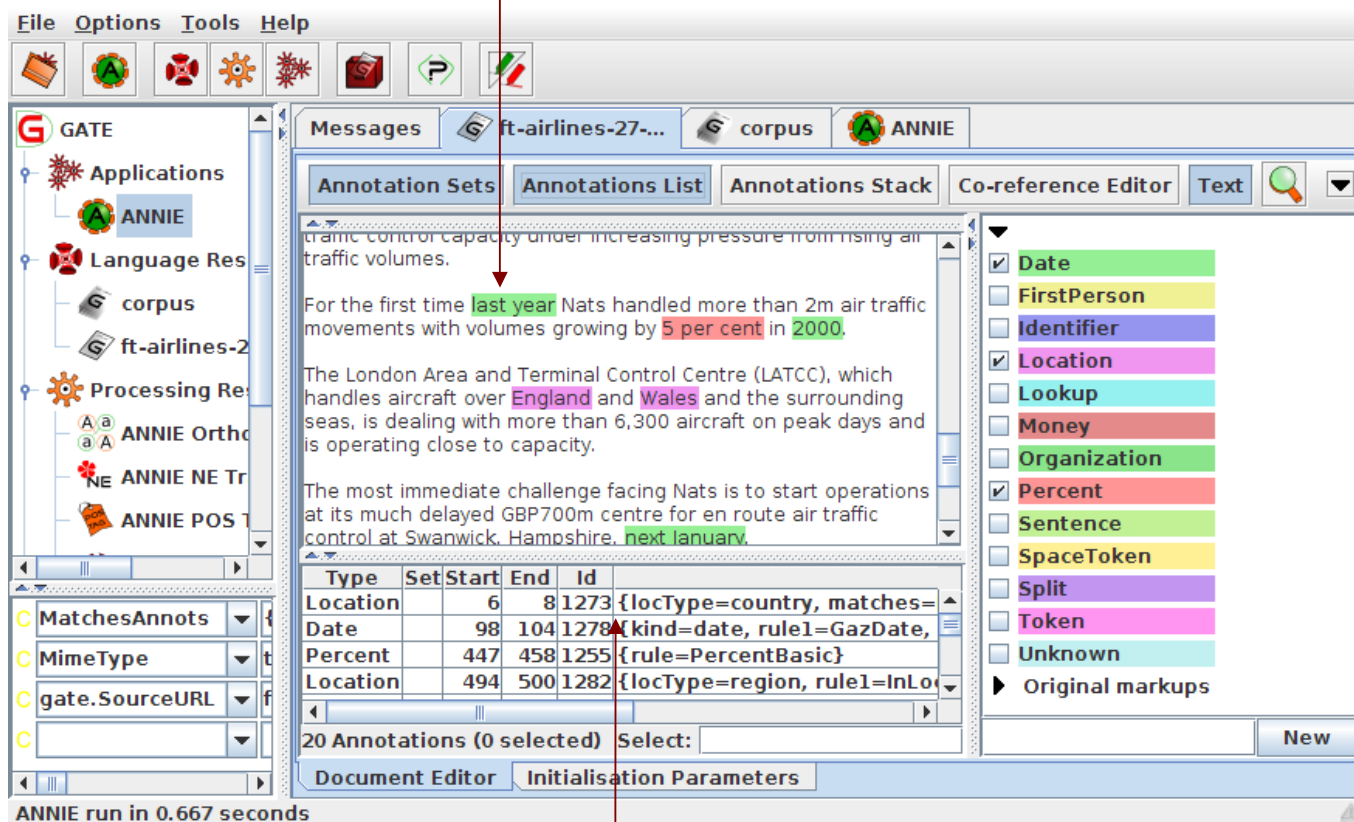- ➢ Select an annotation type to view all the annotations of that type in the document

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
- There are also other annotation views possible such as the AnnotationStack and Coreference Editor: we'll look at these later

Annotations

Date annotation

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

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.

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

➢ Click the edit button [add icon] and add your document to the corpus

- As with the documents, you can name your corpus or use the default GATE name.

➢ Double click on the corpus name in the Resources pane to view the corpus.

➢ Double click the document listed there to view it.

Another way to add documents to a corpus

- You can also create an empty corpus and then add documents to it, if these documents are already loaded in GATE

➢ Create another corpus as before, but do not select any documents to add to it

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

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

  - 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

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 the corpus

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 corpus as before, but 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 in which your documents are stored

- The Extensions parameter lets you select only documents of a certain type.

  - Type "xml" in the box (without the quotes)

- "Encoding" lets you choose the right encoding for the documents. The wrong encoding can cause characters to be incorrectly displayed

  - Enter "UTF-8" here

- "Recurse directories will also load documents in any subdirectories

  - Deselect the "Recurse directories" box

- As if by magic, all the documents will be loaded in one go

- View the contents of the corpus as before.

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 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!

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 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

- ➢ Click the 🅟 icon on the top GATE menu to open the Plugin Manager [or go via File->Manage CREOLE Plugins]

- • A list of plugins will appear

- ➢ 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 "OK"

- ➢ Right click on Processing Resources to see which new PRs are now available

6. Applications

- Loading and running ANNIE and pre-existing applications

- Creating a new application

Here's one I made earlier: ANNIE

- ANNIE is a readymade collection of PRs that performs IE on unstructured text. For those who grew up in the UK, you can think of it as a Blue Peter-style "here's one we made earlier".

- A detailed explanation of ANNIE will be given in Module 2. 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 → Load ANNIE system

➢ Select "with defaults"

➢ Load any document from the hands-on material and add it to a corpus

Running an application

- View the ANNIE application by double clicking on it

- The "Selected Processing resources" table shows the list of PRs in the application, in the order in which they will be run

- Running order is important, since some PRs depend on the results of other PRs

- Click on a PR in the table to see the runtime parameters

- Check that the Corpus box shows a corpus. If it shows "None" then select your corpus using the arrow

    - The corpus is a special runtime parameter that is set for all the PRs in the application

- Click "Run this application"

- This will run ANNIE on the corpus you have selected

Viewing the results

- When a message appears in the bottom left corner of your GATE window saying something like "ANNIE run in 1.3 seconds", the application has 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 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".

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. We will not look at Lucene (searchable) datastores today.

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 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

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 to the datastore

➢ Now populate your corpus (right click on corpus → Populate)

- You should see the documents appear in your datastore

- As if by magic, 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 some special GATE-specific format

- If you want to use your documents outside GATE, you can save them in 2 ways:

    - as standoff markup, in a special GATE representation

    - as inline annotations (preserving the original format)

- Both formats are XML-based. However "save as xml" refers to the first option, while "save preserving format" refers to the second option.

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 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!

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 on a document and select "Save preserving format"

- If the Advanced Option in GATE "Include annotation features for save preserving format" has been checked, then features will be saved as well as annotations, in this mode.

➢ You can play with this option on your own later.

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

- Tomorrow we'll look at the topic of Information Extraction, and 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

➢ Run an application over documents in a datastore