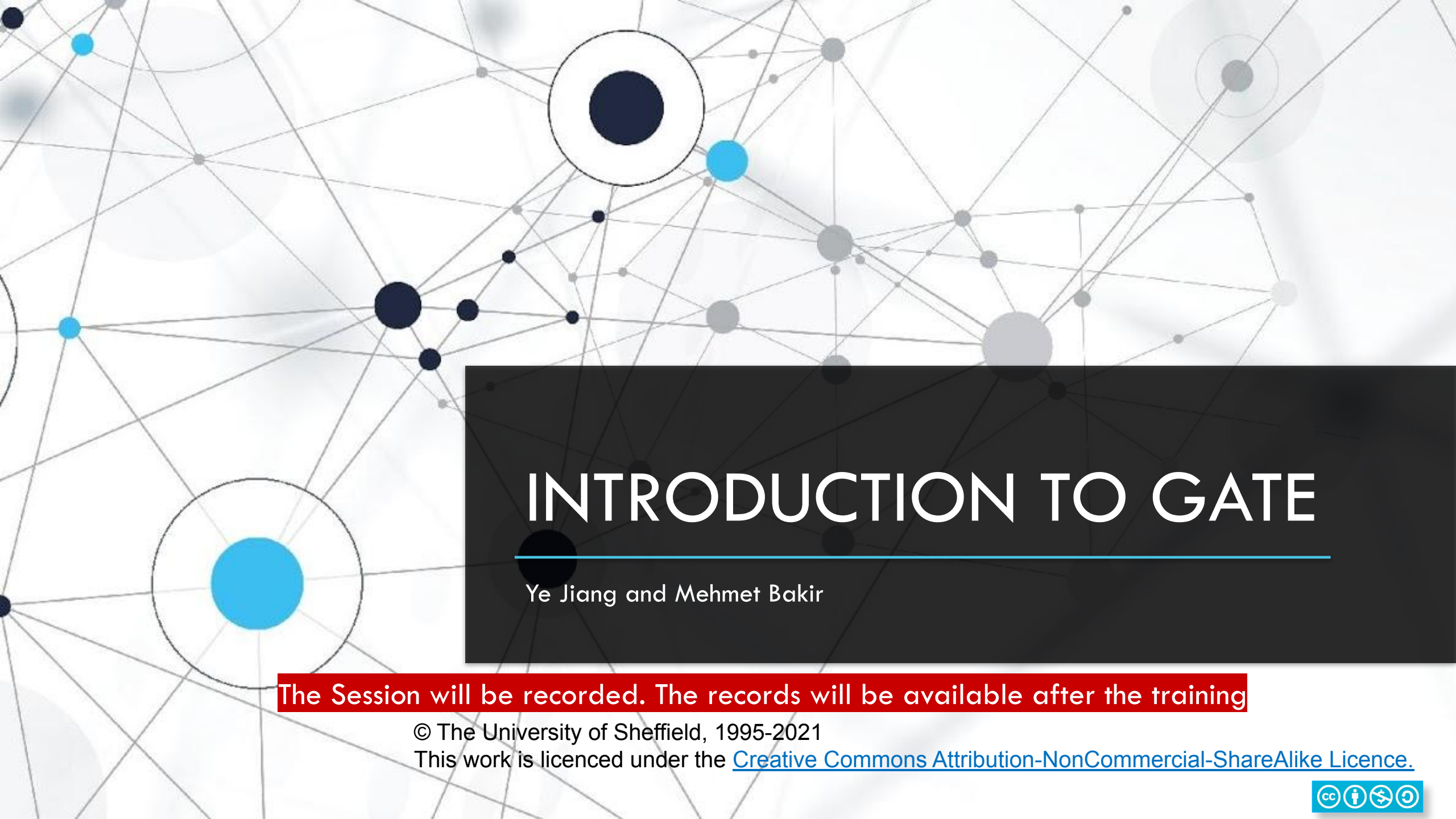


INTRODUCTION TO GATE

Ye Jiang and Mehmet Bakir

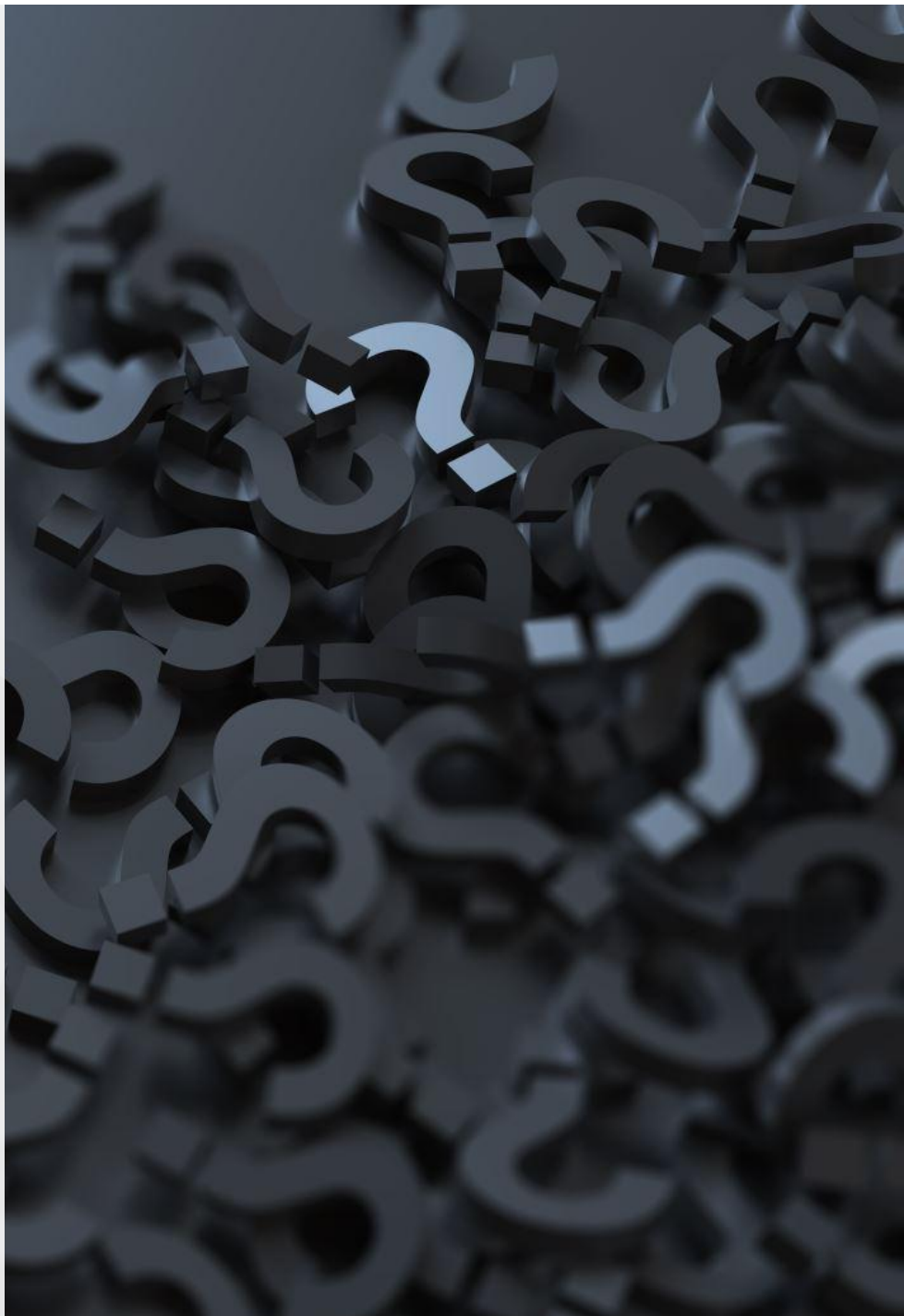
The Session will be recorded. The records will be available after the training

© The University of Sheffield, 1995-2021
This work is licenced under the Creative Commons Attribution-NonCommercial-ShareAlike Licence.

Session 1

1. About GATE
2. Loading and Viewing Document
3. All About Annotation
4. Documents and Corpura
5. Processing Resources and Plugins
6. Applications

*Things for you to try yourself are in red

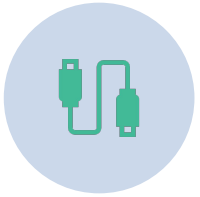


WHAT IS GATE?

GATE is an infrastructure for developing and deploying software components that process Natural Language.

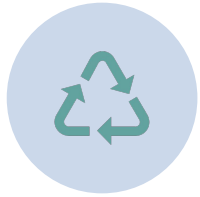
OPEN-SOURCE SOFTWARE FRAMEWORK AND SET OF READY SOLUTIONS FOR TEXT/NATURAL LANGUAGE PROCESSING

A GRAPHICAL USER INTERFACE TO INTERACTIVELY DEVELOP SOLUTIONS (GATE GUI, GATE DEVELOPER)



AN INFRASTRUCTURE OF **PLUGGABLE COMPONENTS** (GATE PLUGINS)

READY-MADE SOLUTIONS TO GET YOU STARTED



RE-USABLE ABSTRACTIONS FOR DOCUMENTS, FORMAT CONVERSION, CORPORA, ANNOTATIONS, STORAGE, ALGORITHMS, …

A (JAVA) LIBRARY PROVIDING A **PROGRAMMING API** FOR USING THE ABSTRACTIONS

COMPANION SOFTWARE FOR SEMANTIC SEARCH (MIMIR)

SCALABLE FROM LAPTOP TO MASSIVE PROCESSING ON THE CLOUD (INCLUDING REAL-TIME STREAM PROCESSING)

GATE 9.0

In this course, we will use **GATE 9.0.**

If you have an older version, please upgrade to the newer one, or many things may not work.

Download (https://gate.ac.uk/wiki/TrainingCourseFeb2021/) and start GATE on your computer (if you haven't already)

TIME TO GET YOUR HANDS DIRTY!

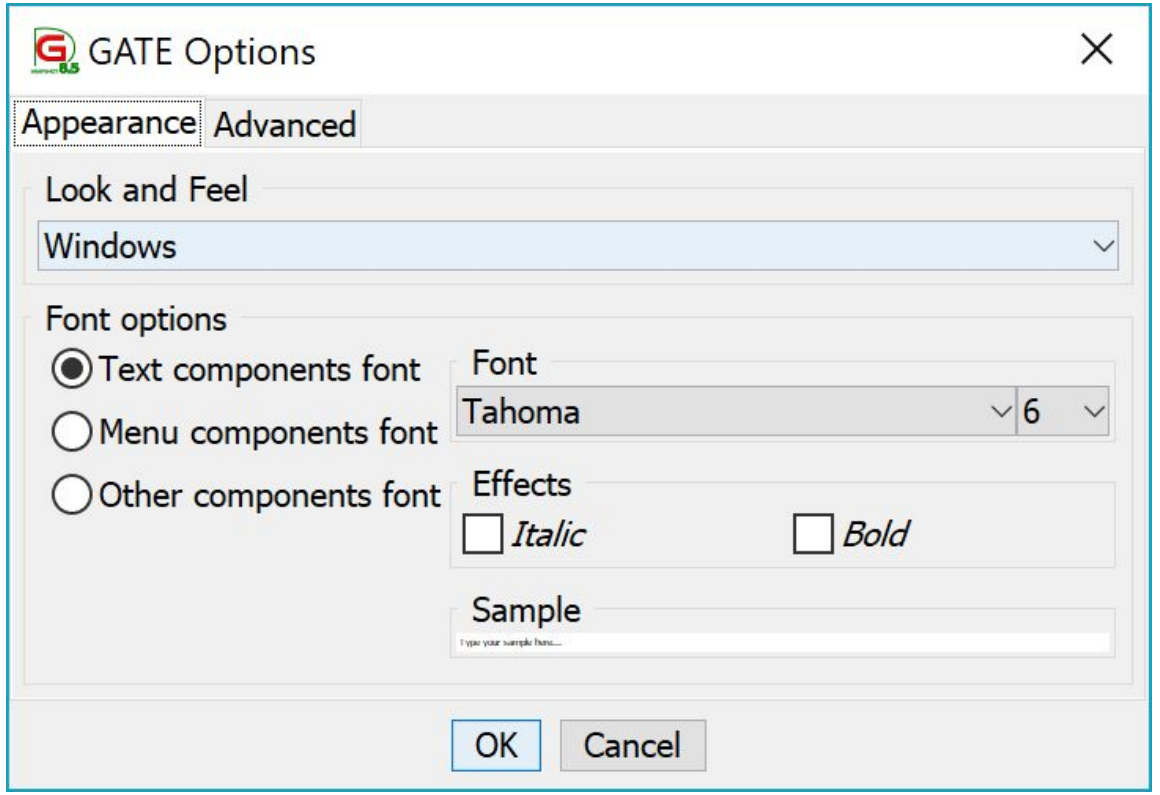
This Photo by Unknown Author is licensed under CC BY-NC-ND

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

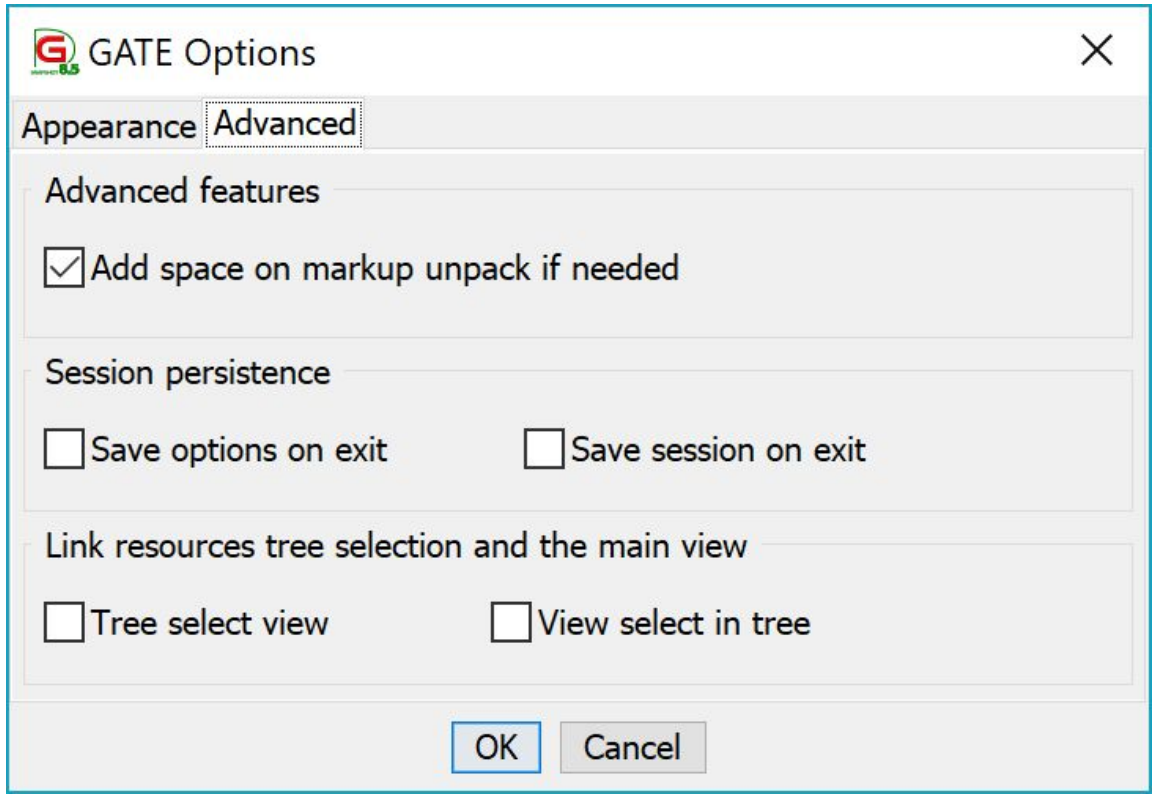
We recommend the Metal Look and Feel (depending on your OS, some features may not work with others)



SETTING UP GATE OPTIONS (2)

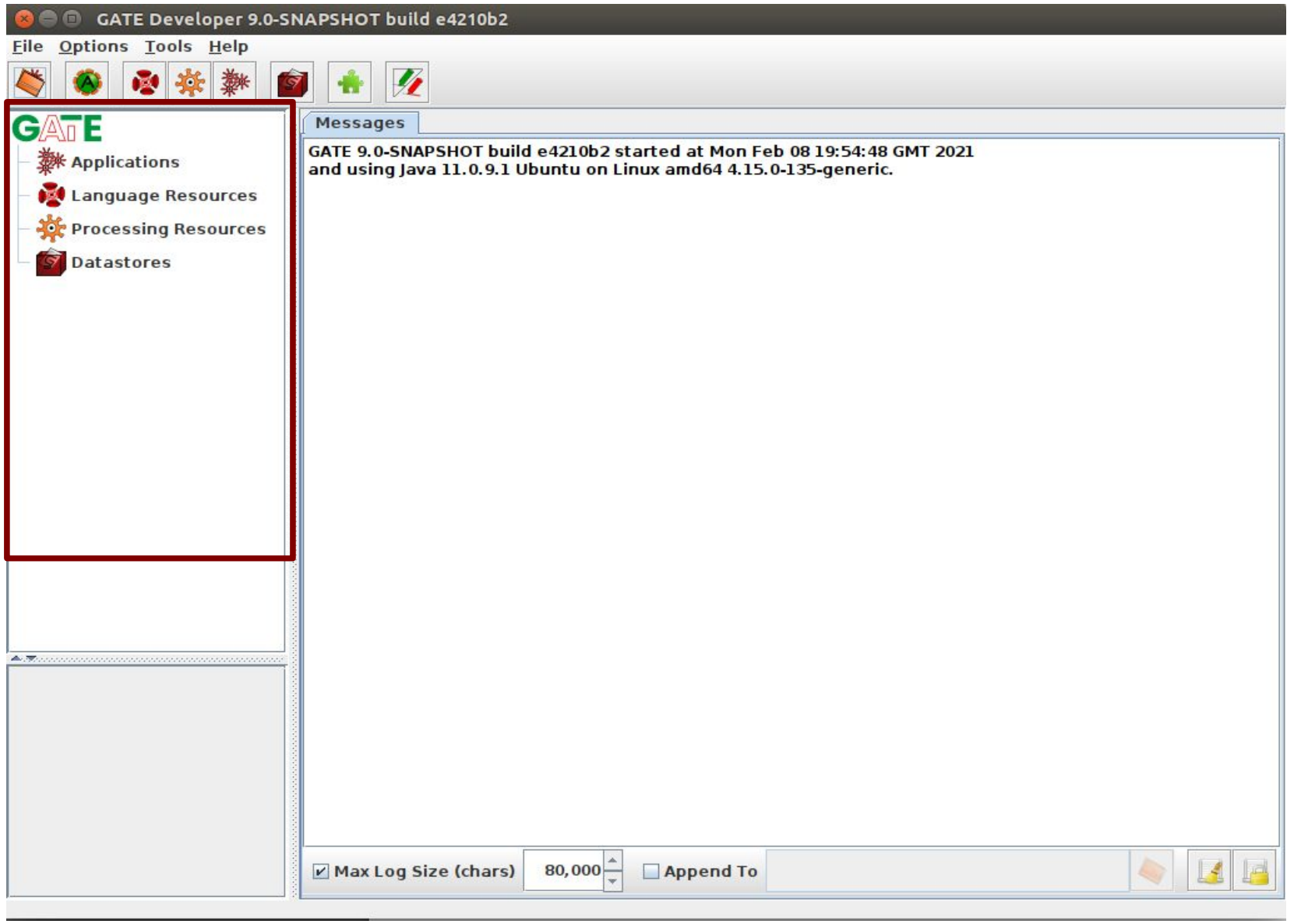
Clicking the **Advanced** tab enables you to adjust settings such as saving your options, and saving the session so that when you reopen GATE, it will remember and reload the applications you had open at the end of your previous session.

You can try this out later.

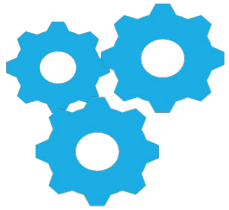


1. FINDING YOUR WAY AROUND THE GUI

Resources Pane



RESOURCES PANE



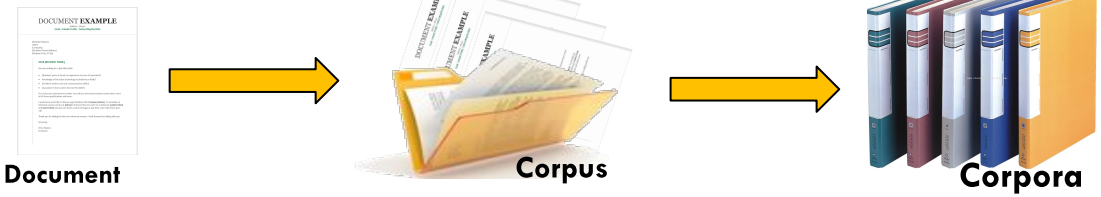
Applications are groups of processes that run on one or more documents

Language resources (LRs) are documents or document collections

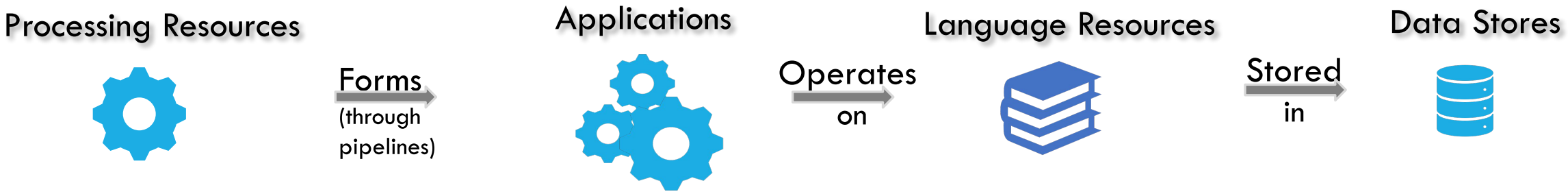
- a collection of documents is known as a **corpus**

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



RESOURCES PANE



SIMPLE OPERATIONS ON RESOURCES

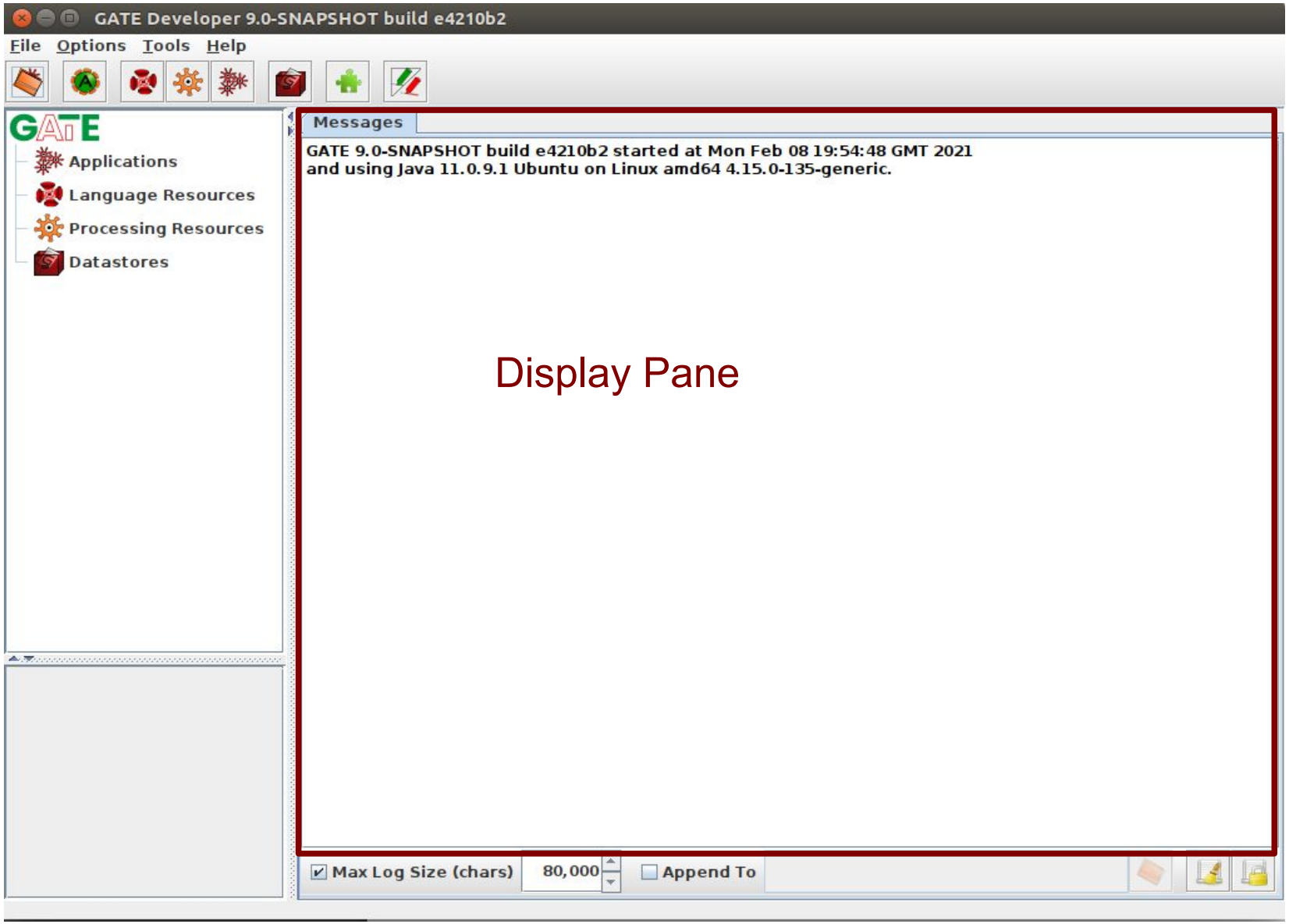
In general, **right clicking** on the name of a resource in the resource pane gives **access to a menu of actions**

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**

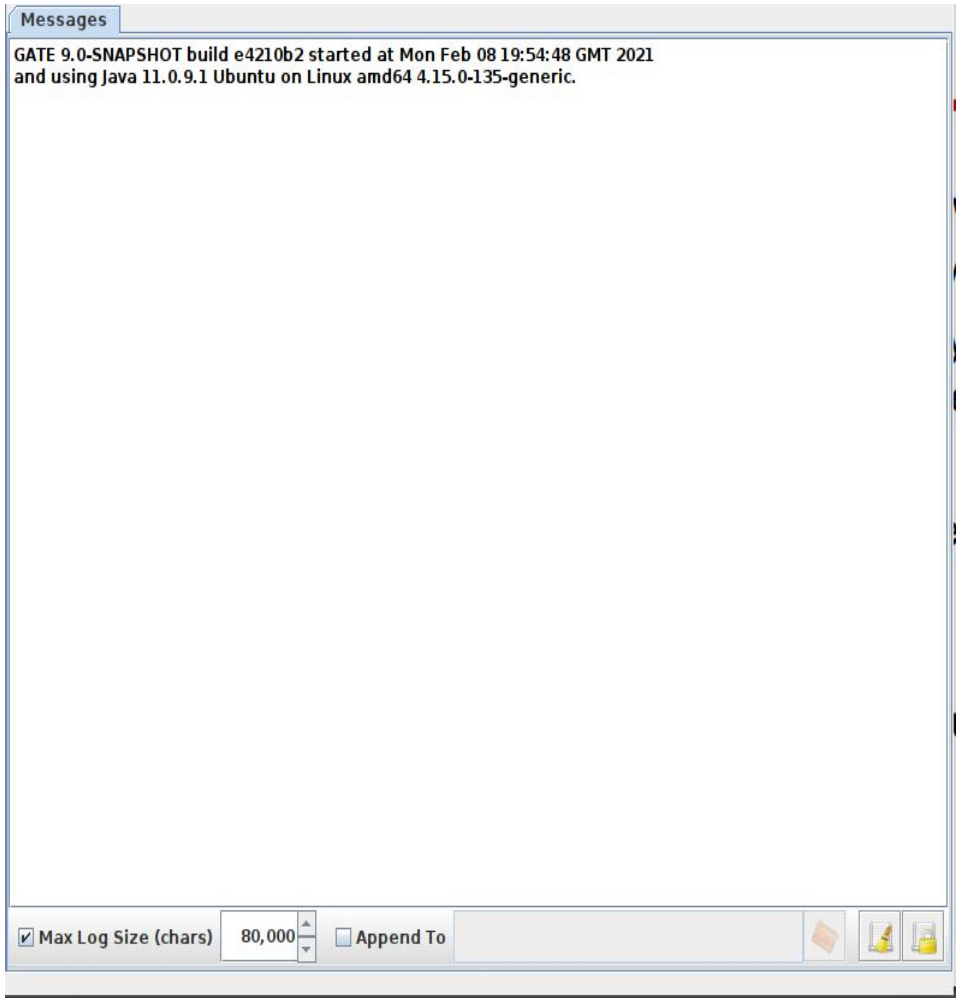
You can also **right click** and then select "**Close**"

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

Parameters enable different settings to be used.

Applications, LRs, and PRs all have various **parameters which can be set either at load time (initialisation) or at run time.**

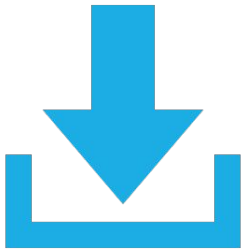
- **Initialisation Parameters** (set at load time) cannot be changed without reloading.
- **Run time Parameters** can be changed between each application run.

2. LOADING AND VIEWING DOCUMENTS

Loading a document and setting its parameters

Navigating through documents and viewing their annotations

LOADING A DOCUMENT

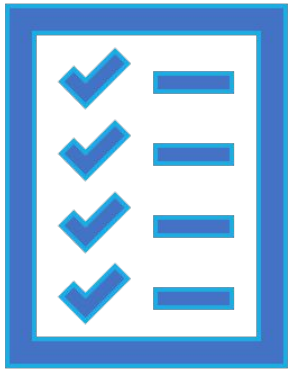


- When GATE loads a document, it converts it into a special format for processing.
- GATE can process documents in **almost all kinds of formats**: e.g. plain text, HTML, XML, PDF, Word.
- Documents have a **markupAware** parameter which is set to true by default: this ensures GATE will **process any 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.
  - 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 file from your annie hands on materials **and** one from the Web – you must include http:// when specifying a URL

INITIALISATION PARAMETERS

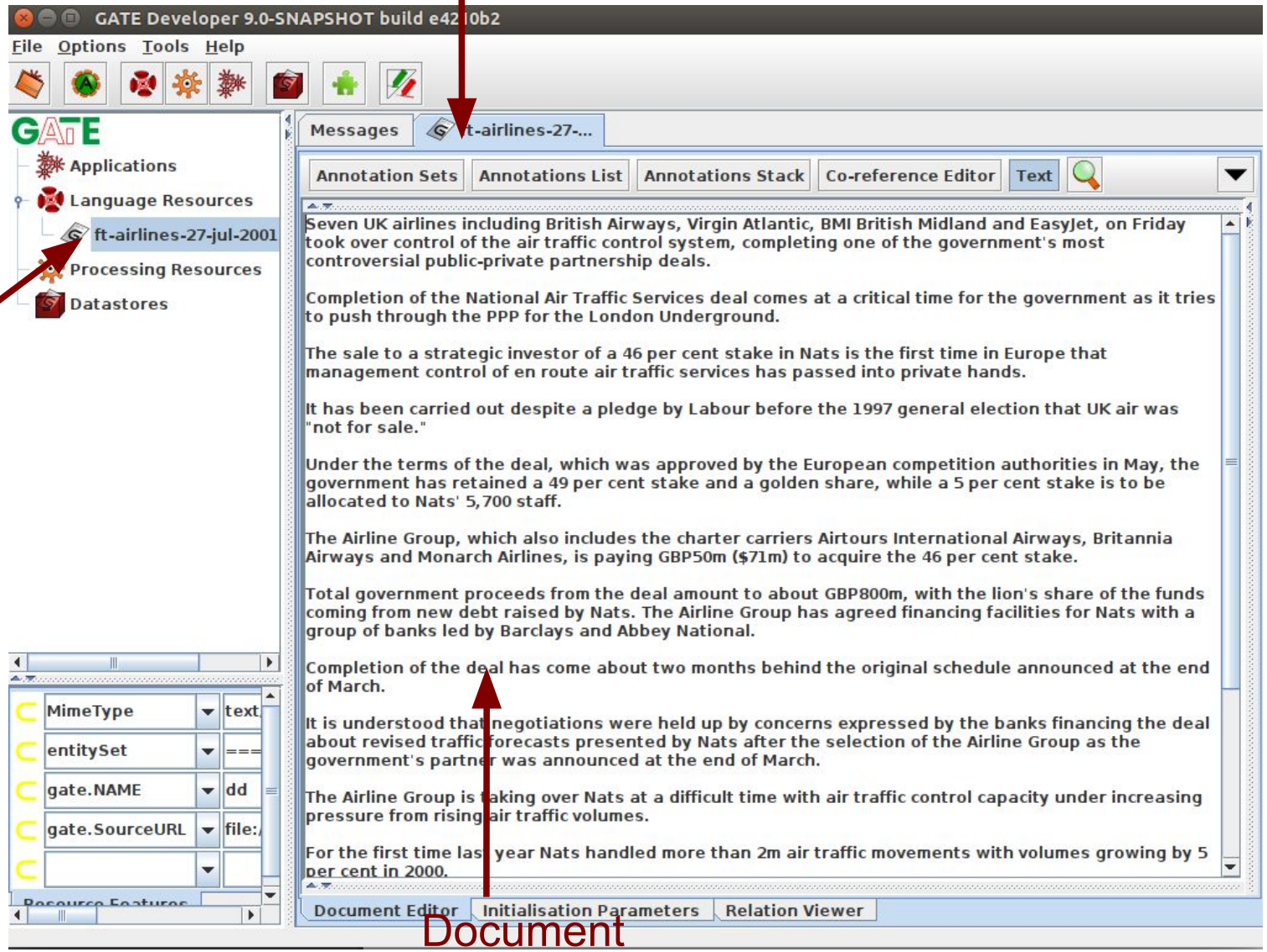


- A document has a variety of **init** parameters: some compulsory and some optional.
- Compulsory parameters have **a tick** in the "**Required**" box.
- You can provide your own name or use the default name GATE provides (document name + a unique ID, which prevents confusion with multiple copies of 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 tabs

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 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 are metadata associated with a document segment.

“The main **purpose of GATE is annotating documents**. Whilst applications can be used to annotate the documents entirely **automatically**, annotation can also be done **manually**, e.g. by the user, or **semi-automatically**, by running an application over the corpus and then correcting/adding new annotations manually.”

Each annotation consists of;

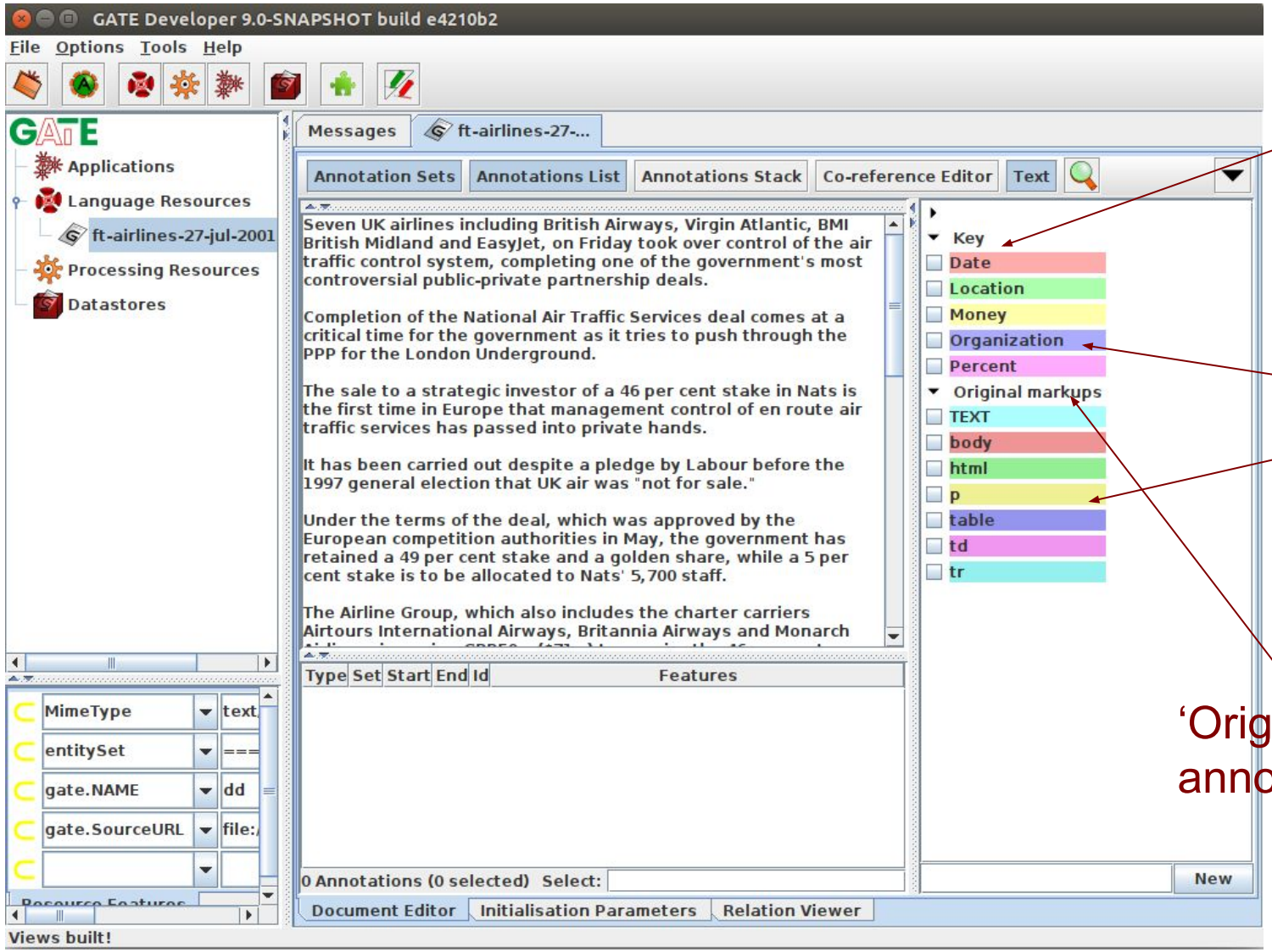
- start and end offsets
- optionally a **set of features associated with** it
- each feature has a **name and a value**

Annotations are created, deleted and managed by **annotation sets.**

ANNOTATION SETS

- Annotations are **grouped into sets**, e.g. Default, Original Markups
- Each set can contain several **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**

ANNOTATION SETS



'Key' 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)

  Key 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

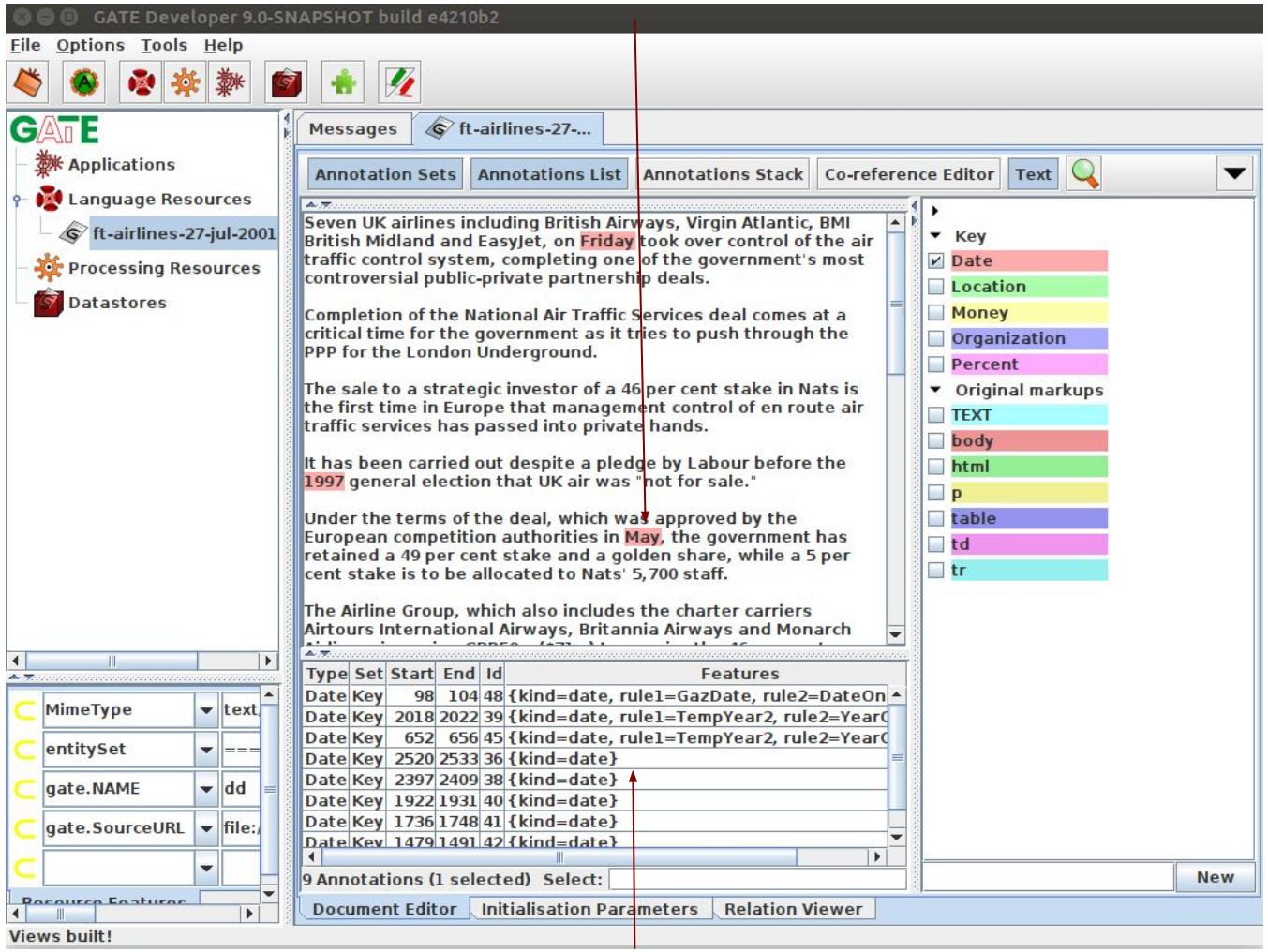
  Table shows annotation type, annotation set, offsets, features that includes names and values

  Select a row in the table to highlight the annotation in the text

- Click on a column heading to sort according to the header

ANNOTATIONS

Date annotation



Annotations list

EDITING EXISTING ANNOTATIONS

- **Select an annotation type from the Annotation Sets** view and hover over a highlighted annotation in the text

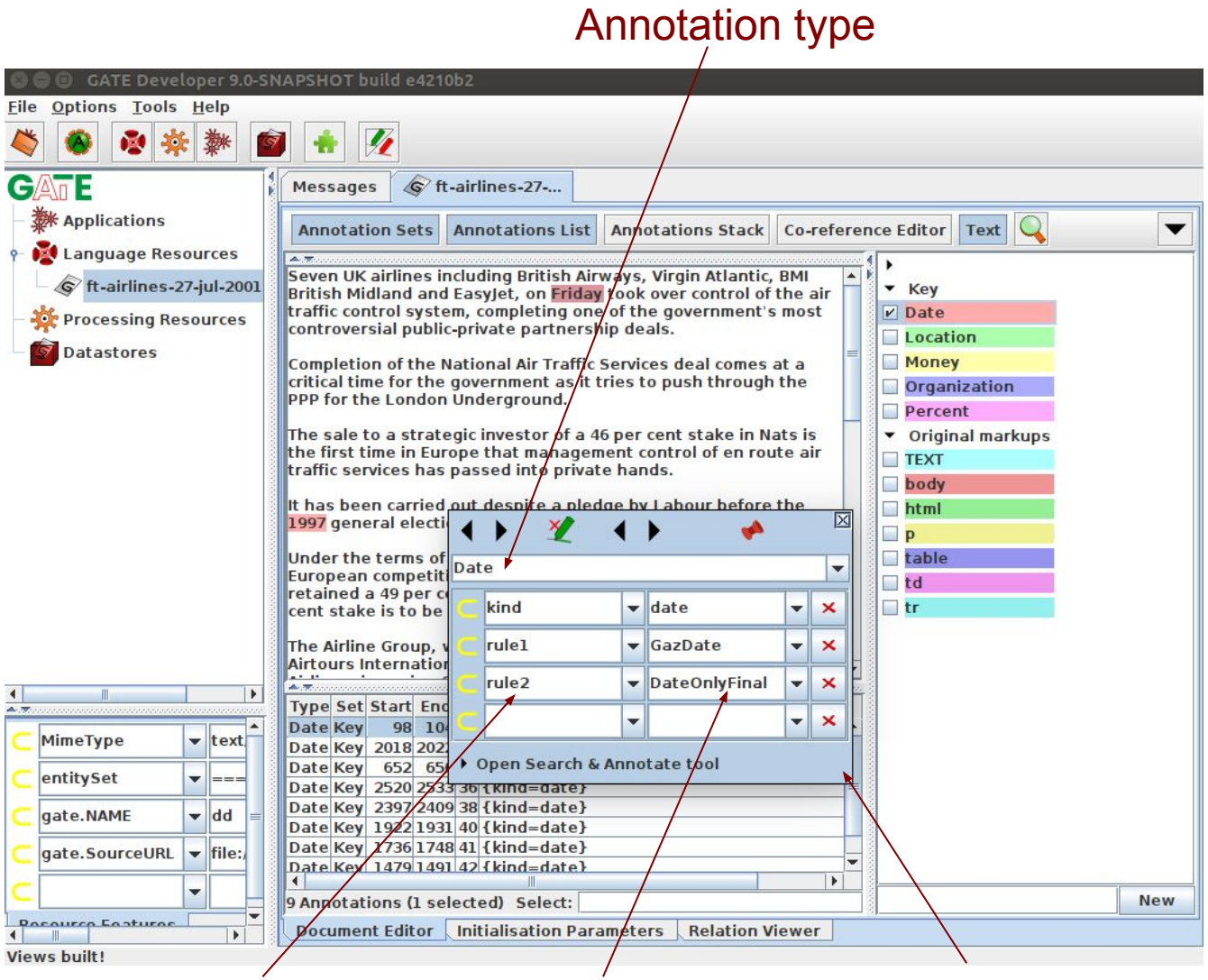
  A popup window displays more information about it: **this is the annotation editor**

- **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



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

Please close all open documents and corpora

- **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"**
- Leave "Mime type" as blank.
- "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 CORPUS

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.

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.

An application consists of any number of PRs

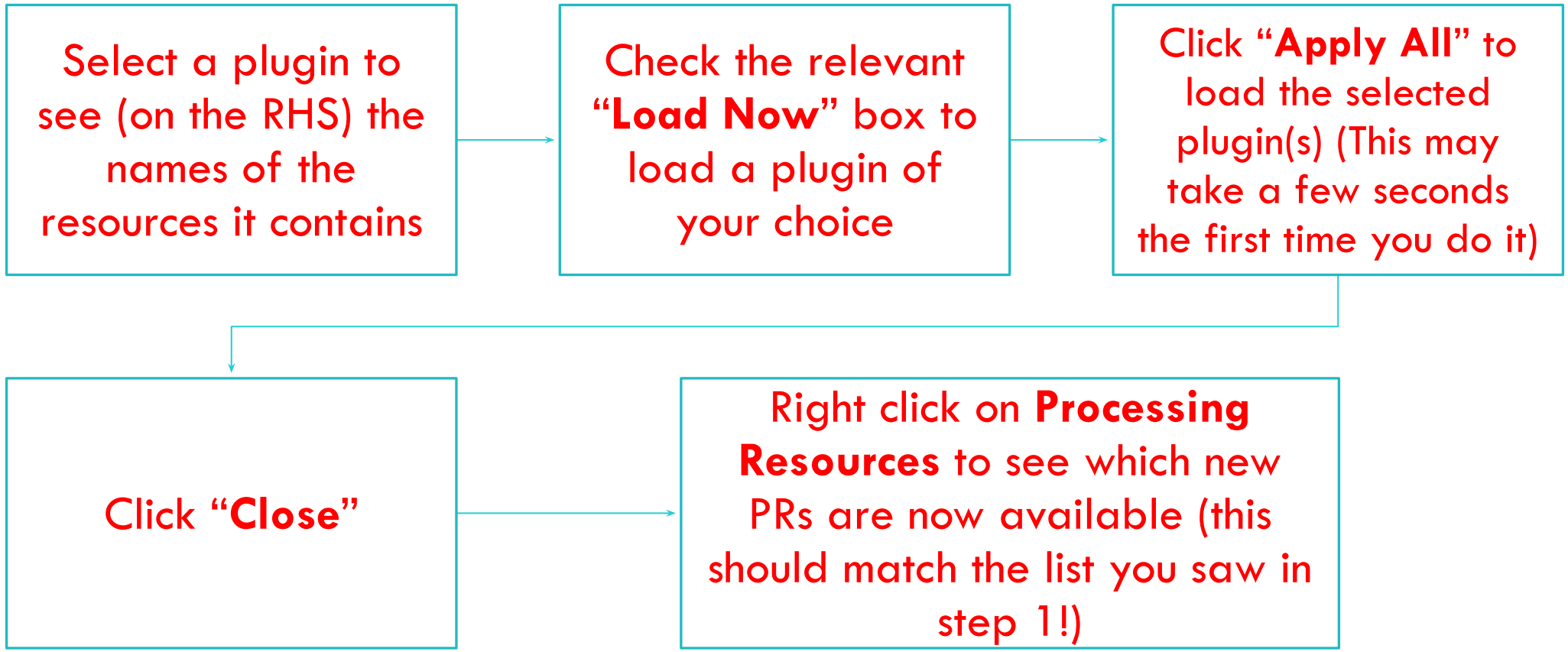
A plugin is a collection of one or more PRs

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]
- 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

PLUGIN MANAGER



DOWNLOADING THE RESOURCES FROM A PLUGIN

- 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.**

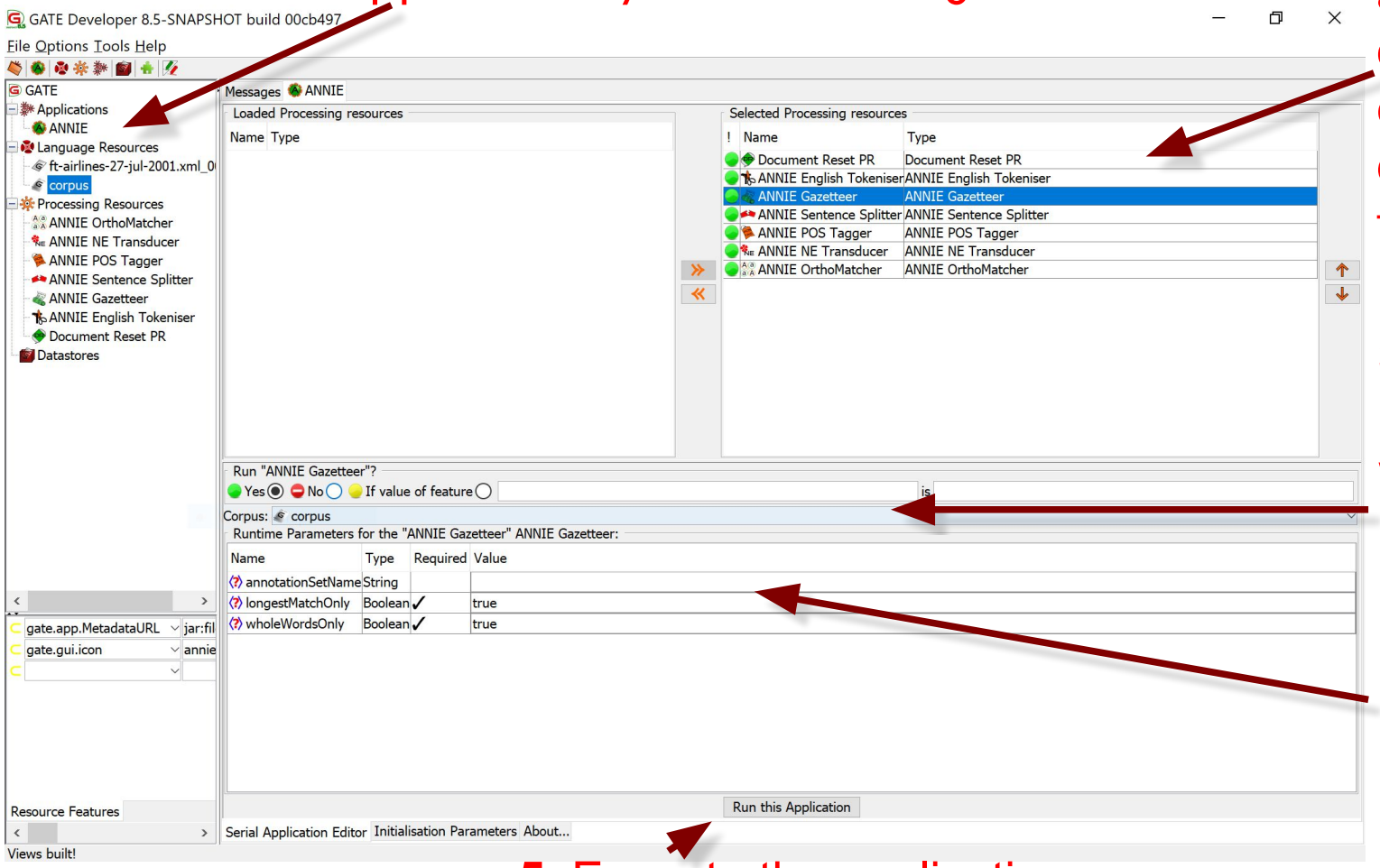
- **Click the** **icon** from the top GATE menu **OR** Select

  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

- **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.

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

Change the name of annotation set to: ANNIEresult

- **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)

- Double click on ANNIE.
- **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.

SUMMARY

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