

BD003: Introduction to NLP

Dr. Diana Maynard

University of Sheffield, UK

Part 1:

Introduction to GATE

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 person who has led the development of the NLP tools in GATE since 2000 is the one presenting to you now 😊
- And by the way, just because it's old doesn't mean it's out of date.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)

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.
- Things suggested for you to try yourself are in **red**.
- **Start GATE on your computer now (if you haven't already) by double clicking the icon**
- Please don't jump ahead: if you're already finished with a task, perhaps you can help your neighbour if they get stuck.
- Please try to keep questions during the sessions related to the current topic
- There will be time at the end or in the breaks for more general questions

GATE GUI

Menu Bar

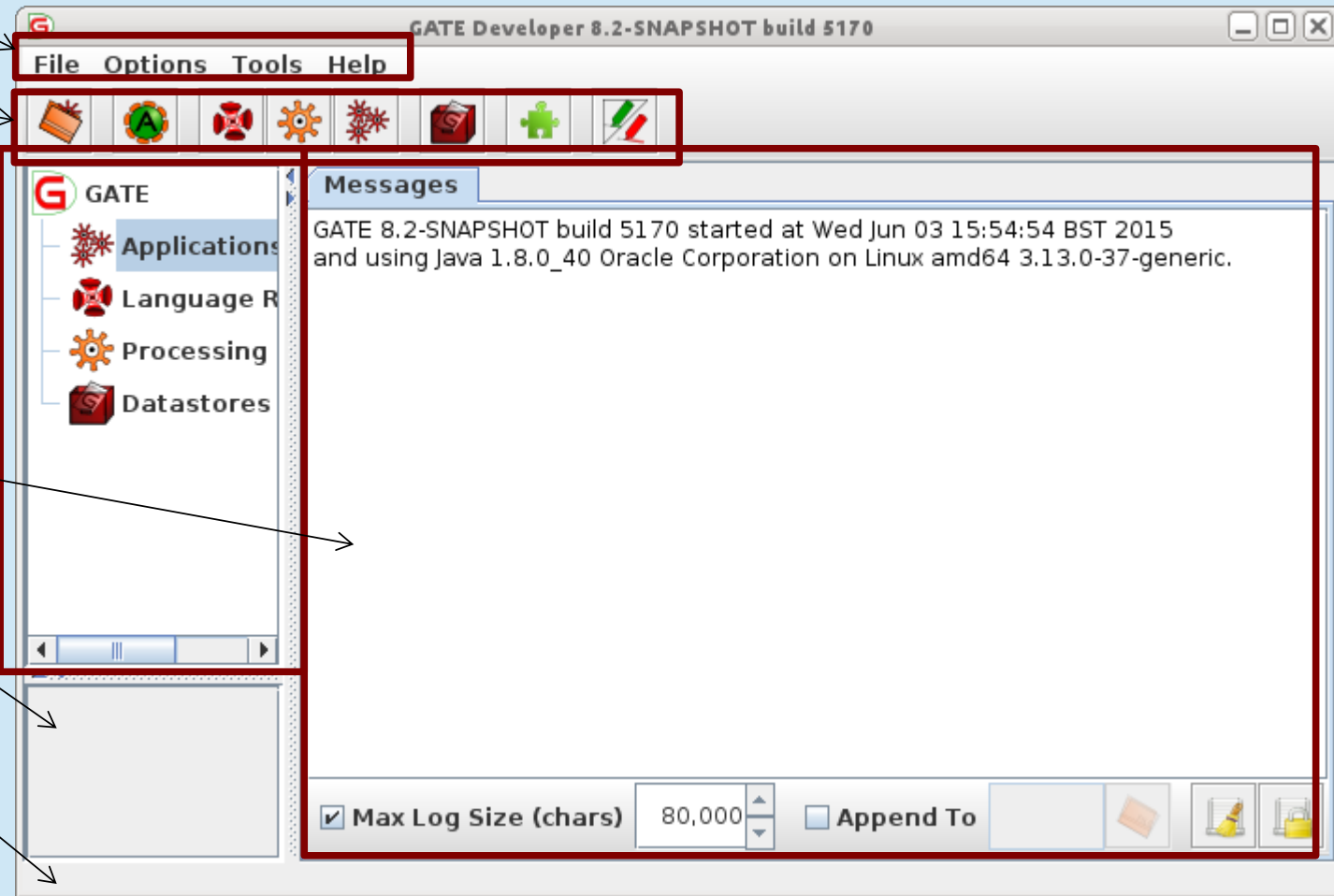
Shortcut
Buttons

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
- **Datastores** are for storing documents and corpora for later use
- **Applications** (“pipelines”) are sequences of processing resources that run on one or more documents

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 in the resources pane will display it
- Tabs along the top of the display pane allow you to choose which of the open resources to display

Create New Document

- **From the Resource Pane, right click “Language Resources” → New → GATE Document**
- Ignore the parameter settings that will be displayed
- **Click OK**
- “GATE Document_<id>” will now be 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

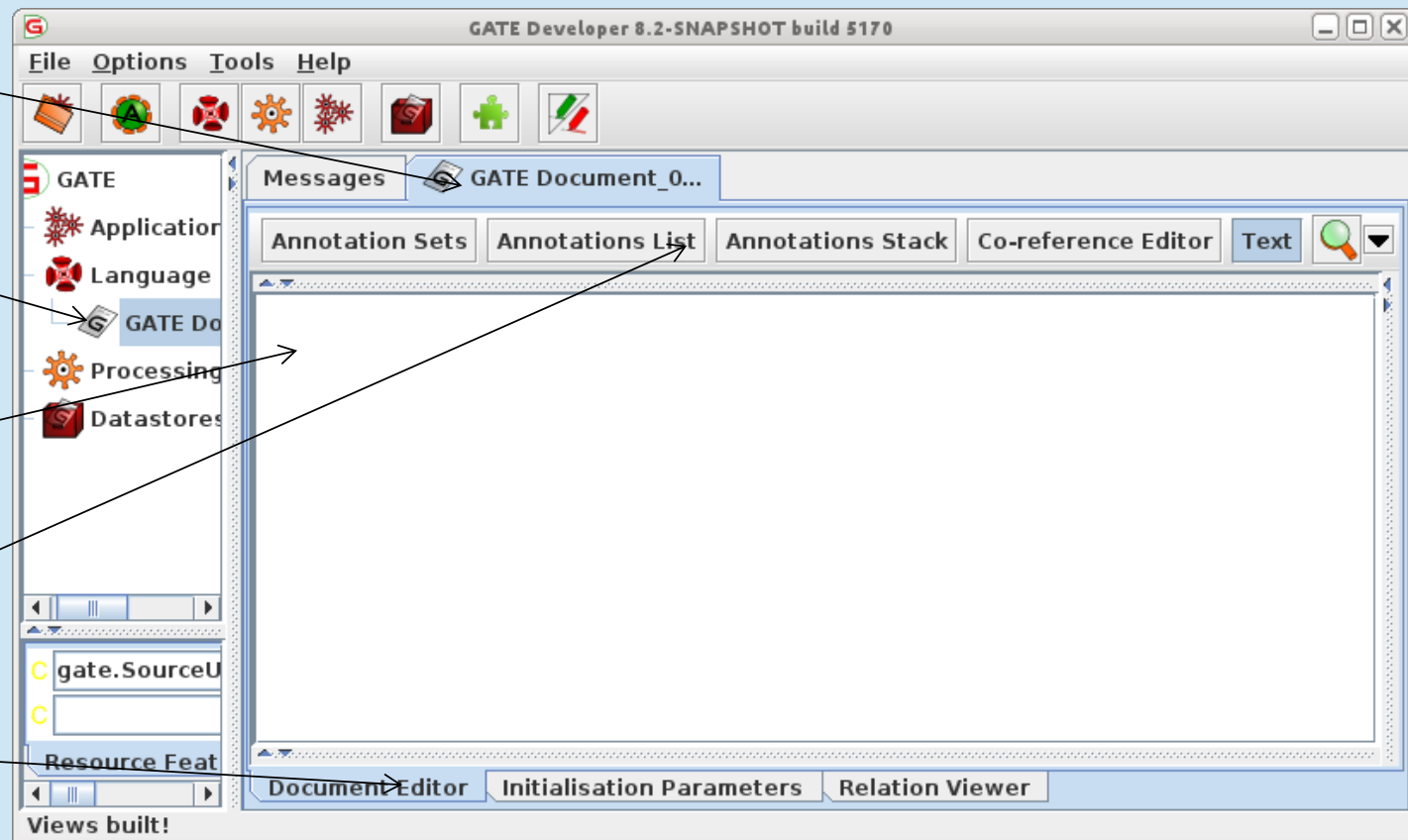
Document
Tab

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)

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+BS) 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 (init) Parameters**
- Processing resources can also have parameters which can be changed for each run: **Runtime Parameters**
- Init parameters specify how a resource is created, e.g. the location of a document to load
- 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 , CSV, JSON
- GATE can load documents from files and from URLs
- When a document is loaded, it gets converted to GATE internal format as document text + annotations.

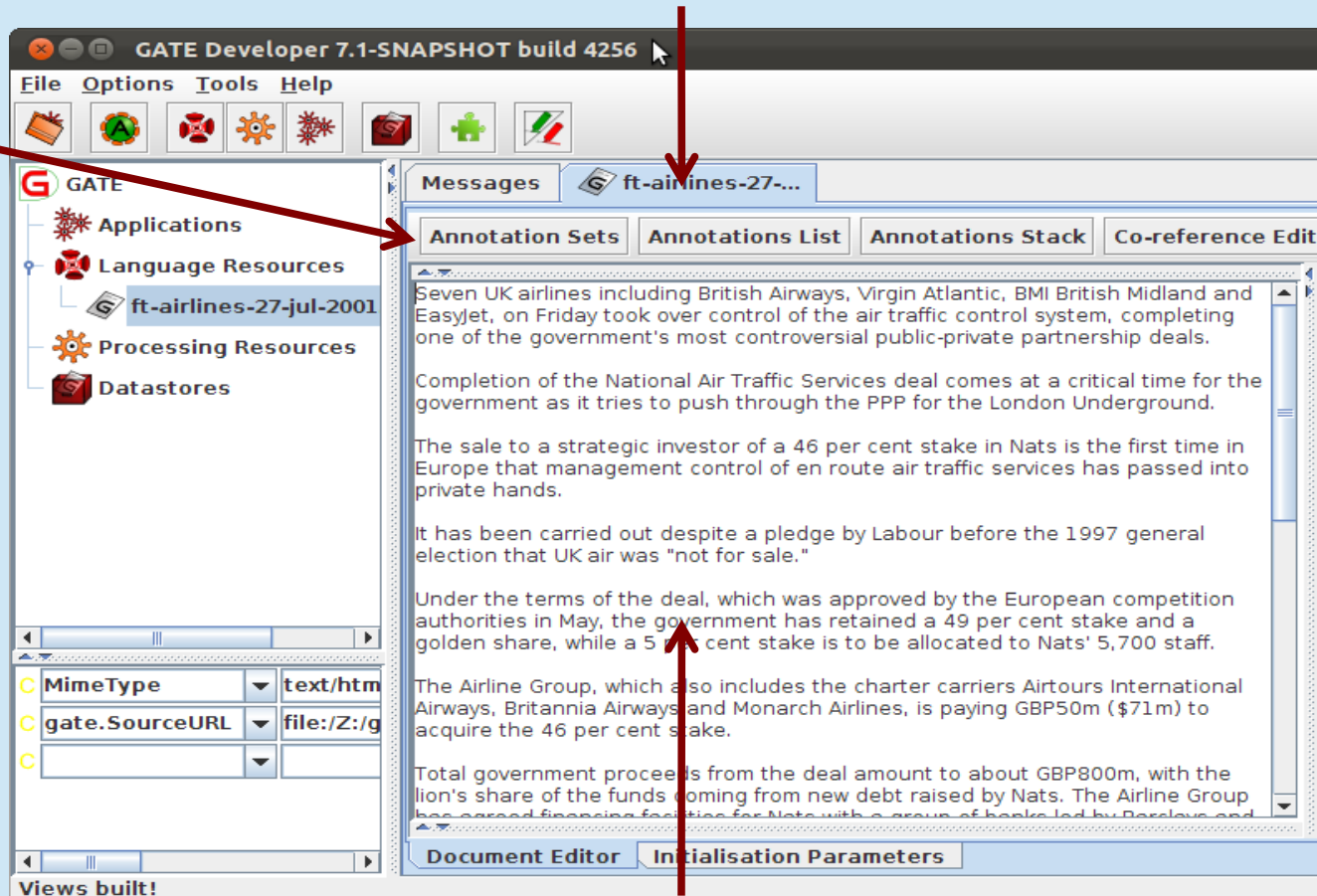
Loading a document

- 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:**
corpora → 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>**
- Note: if you use the BBC page above, we suggest picking a story and clicking on it to get a better document for processing, as the main news page contains mainly just links

Document viewer

Highlighted tab is the resource currently being viewed

Document
viewer
buttons



Document

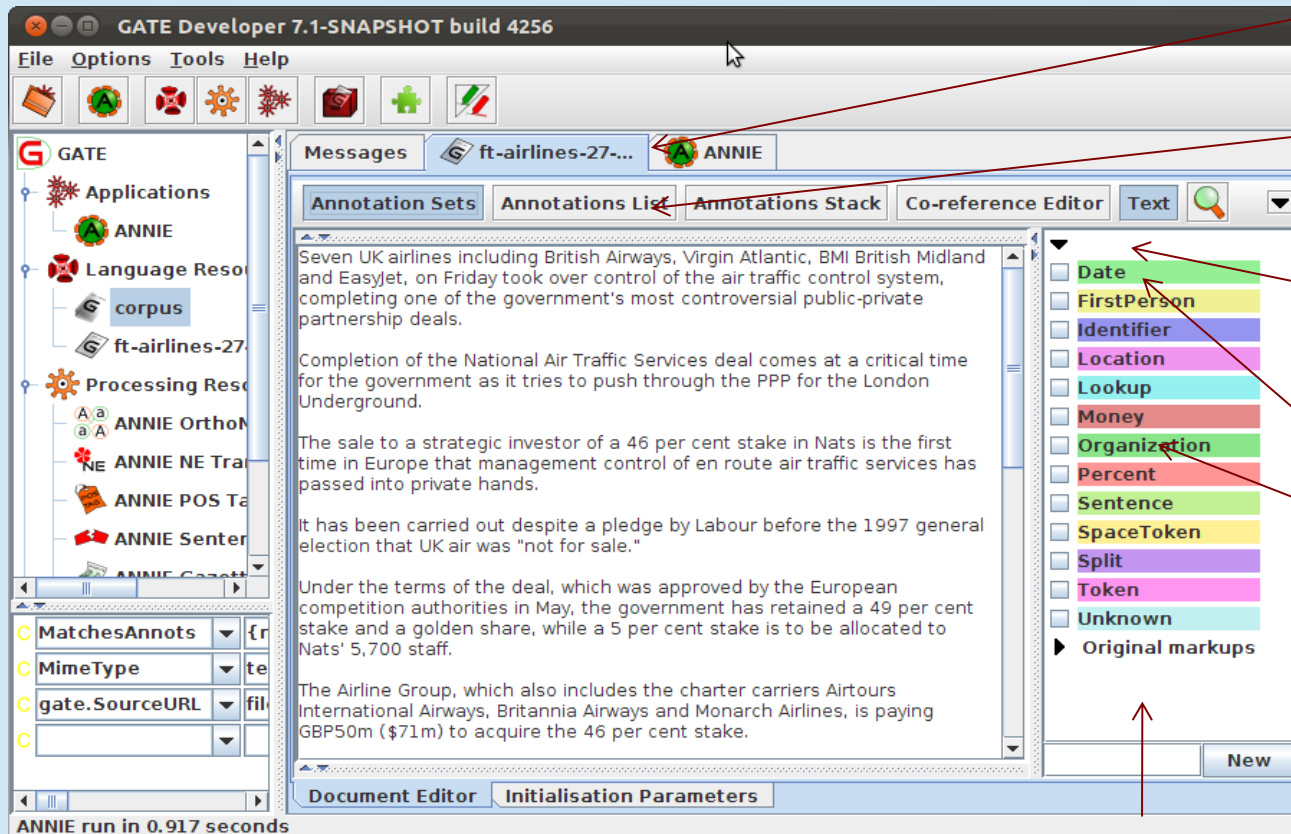
Annotations

- 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”
- 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 gold standard annotations
- 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**
- Table shows annotation type, annotation set, offsets, annotation id, and features (for all selected annotations)
- **Select a row in the table to highlight the annotation in the text**
- There are also other annotation views possible such as the Annotation Stack and Coreference Editor

Annotations

Date annotation

The screenshot shows the GATE Developer 7.1-ANNIE interface. The main window displays a text document with several sentences. Annotations are visible on the text, including "last year", "5 per cent", "2000", "England", "Wales", "next January", and "next 10 years". The "Annotations List" tab is active, showing a table of annotations. The table has columns: Type, Set, Start, End, Id, and Fea. The table contains 10 rows of annotations. The "Date" annotation type is highlighted in the "Annotations List" tab. The "Date" checkbox is checked in the "Annotations List" tab. The "Date" checkbox is checked in the "Annotations List" tab.

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}

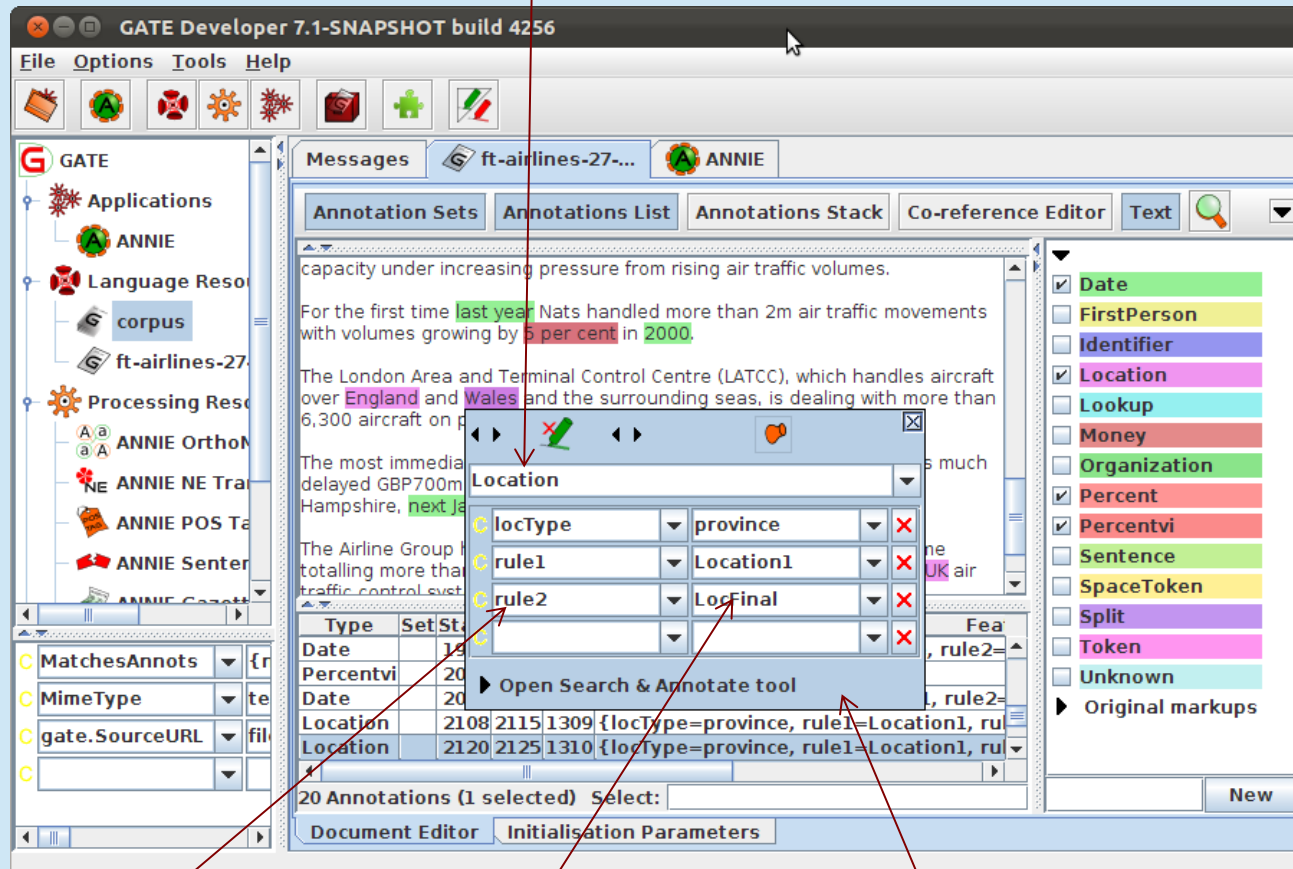
Annotations table

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)
- 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

Annotation type



feature name


value

annotation editor

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 (corpora/news-texts)
- All the documents will be loaded in one go
- View the contents of the corpus as before

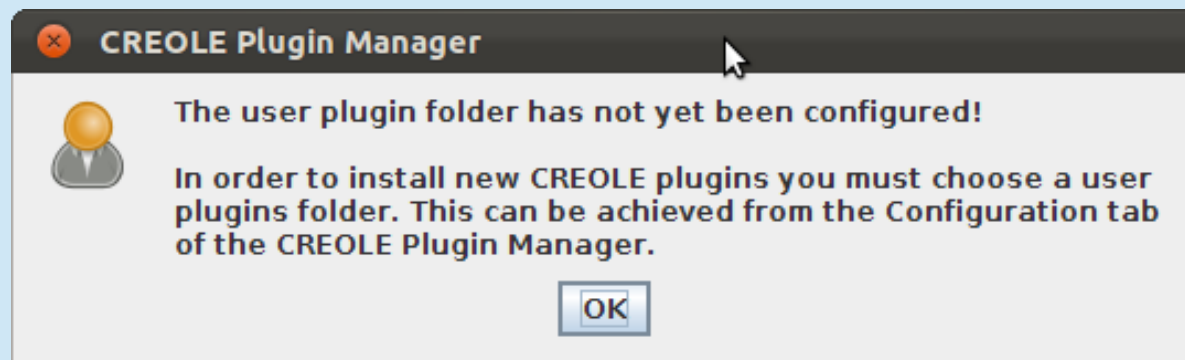
Processing Resources and Plugins

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” or “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 ANNIE is in the ANNIE 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 are loaded via the Plugin Manager (green jigsaw piece icon)

Plugins

- Click the  icon on the top GATE menu to open the Plugin Manager [or go via File → Manage CREOLE Plugins]
- Depending on your version of GATE, you may see a popup box:



- The user plugin folder is a folder on your computer where plugins other than those provided by GATE are stored

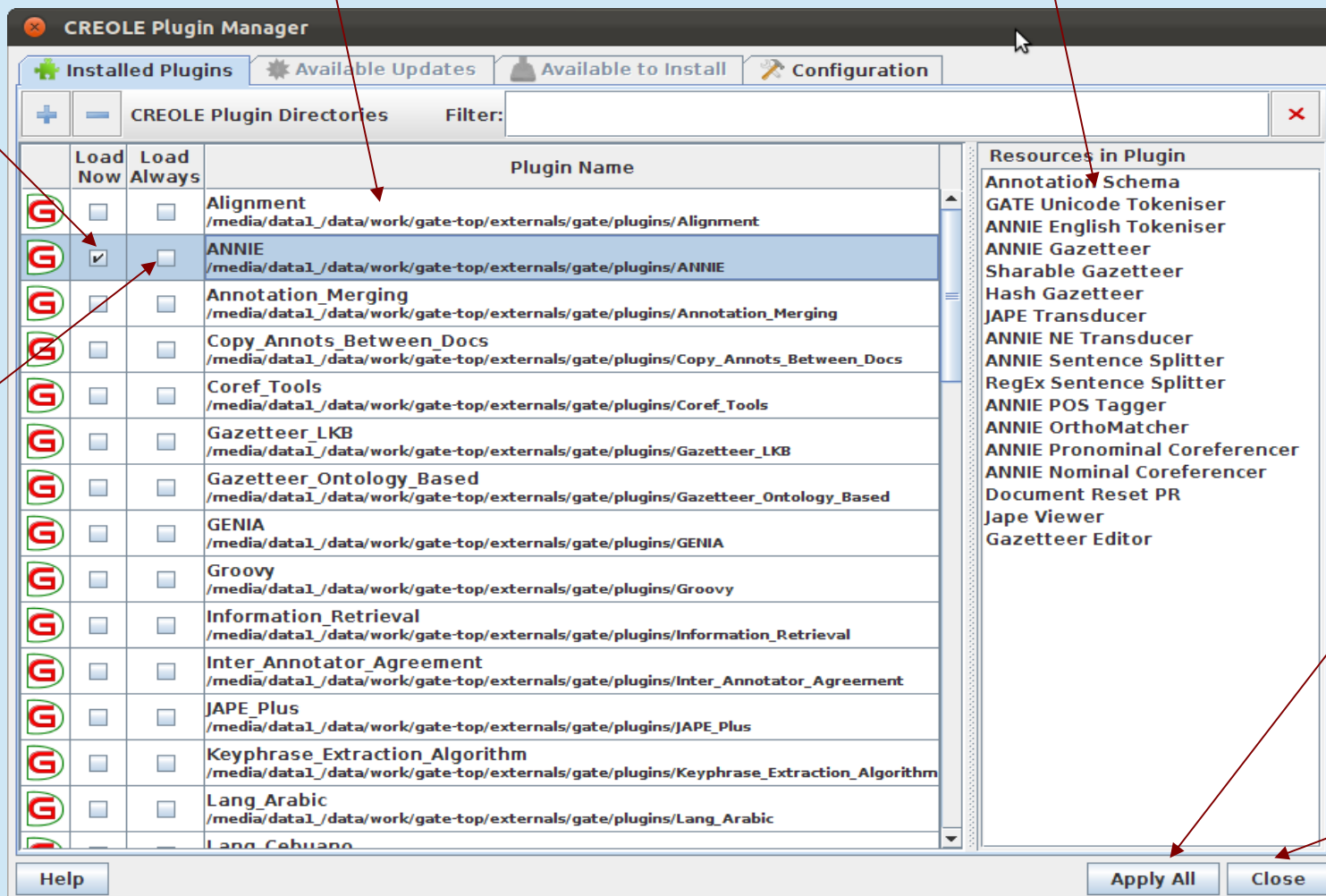
Plugins

Load the plugin for this session only

Load the plugin every time GATE starts

List of available plugins

Resources in the selected plugin



Apply all the settings


Close the plugins manager

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
- Click “Close”
- Right click on Processing Resources to see which new PRs are now available

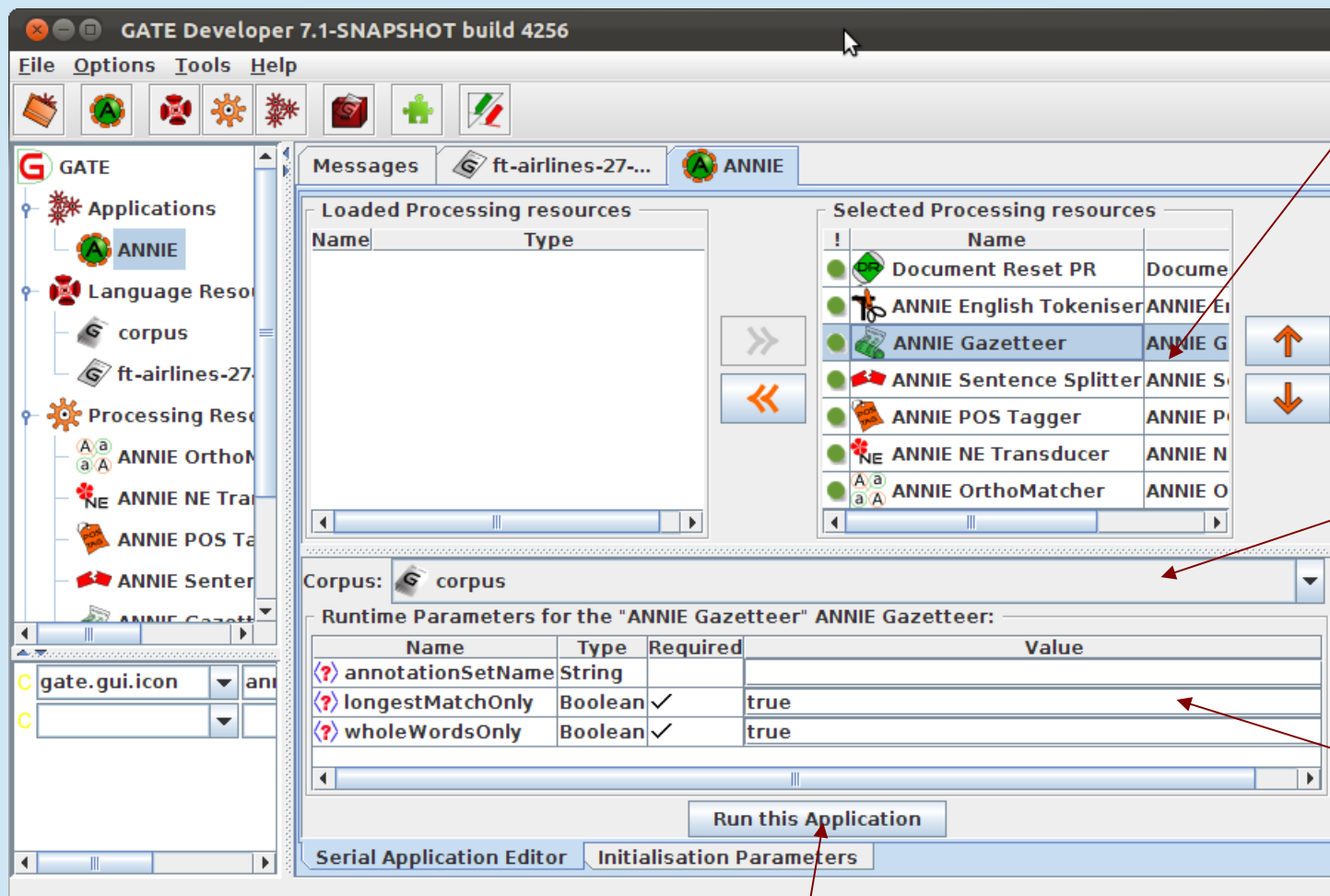
Applications

Here's one I made earlier: ANNIE

- ANNIE is a readymade 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 → 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



PRs selected in application (in order of their execution)

Corpus on which the application is executed

Runtime parameters of the selected PR

Execute the application

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.

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”

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.
- Run the application and view the results on the document.
- You should see a new annotation type “VG”.

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

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

Summary

- This first session 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
- Next, we'll look at various NLP components, and further examine ANNIE, GATE's default Information Extraction system