

# Tutorial on Text Mining for the Going Digital initiative



**Natural Language Processing (NLP), University of Essex**

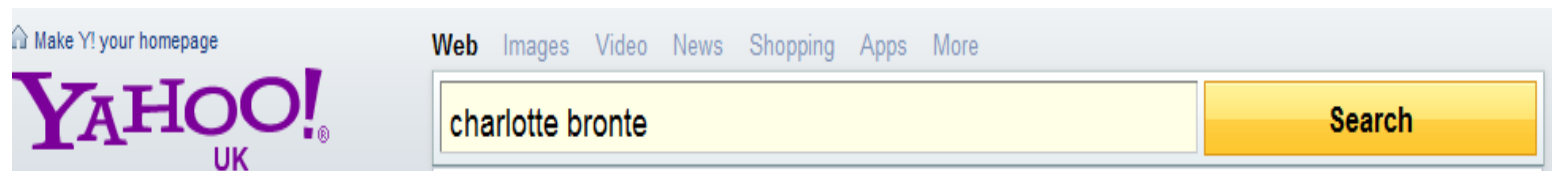
6 February, 2013

# Topics of This Tutorial

- Information Extraction (IE)
- Examples of IE systems
- GATE (General Architecture for Text Engineering)
  - Introduction to GATE developer
  - Guided tour of the GATE GUI
  - Language Resources, Datastores, Applications and Processing Resources
  - Annotations
  - ANNIE Tool , GATE's default IE system

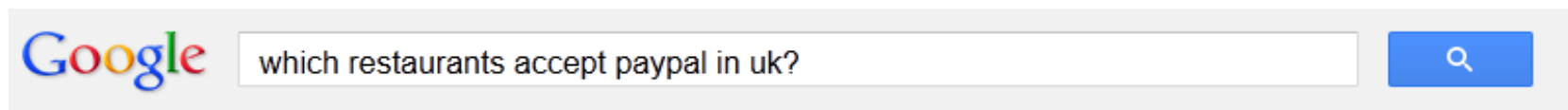
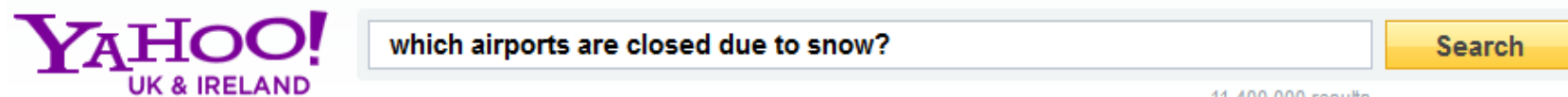
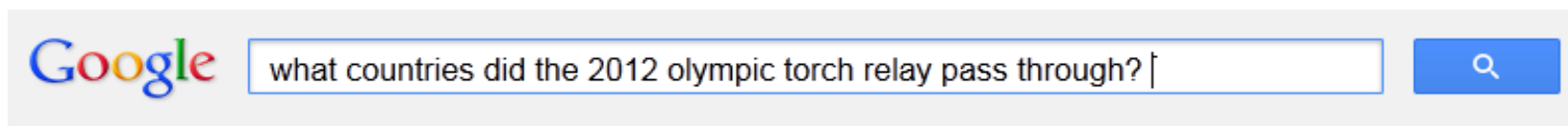
# Information Retrieval vs. Information Extraction

- Information Retrieval
  - large text collections (Web) → Documents



# Information Retrieval vs. Information Extraction

- Getting Facts can be hard and slow.



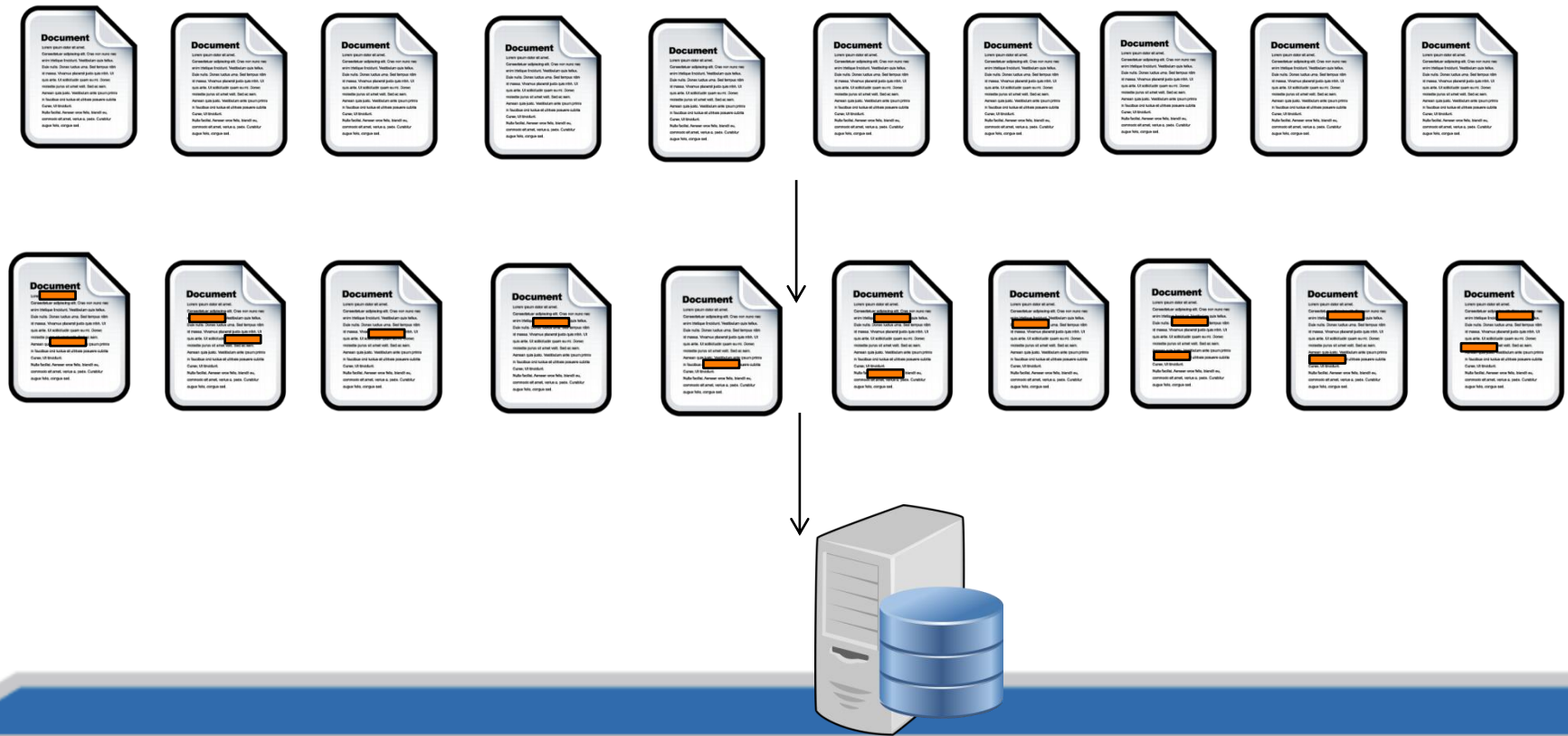
# Information Retrieval vs. Information Extraction

- Information Extraction

➤ large text collections

Facts

Structured Information



# Named Entity Recognition

- Cornerstone of IE
- Identification of proper names in texts,
- Classification them into a set of predefined categories
  - Person
  - Organisation (companies, government organisations, committees, etc)
  - Location (cities, countries, rivers, etc)
  - Date and time expressions
- other types are frequently added, as appropriate to the Application.
  - Emperors
  - Historical events.
  - Ships, etc.

# Named Entity Recognition

Person

Location

Organisation

Year

Mark lives in London. He has worked at University of London since 1990.

# Co-reference

Person

Location

Organisation

Year

Mark lives in London. He has worked at University of London since 1990.





# Relations

Person

Location

Organisation

Year

Mark lives in London.

He has worked at University of London since 1990.



Live\_in

# Relations

Person

Location

Organisation

Year

Mark lives in London. He has worked at University of London since 1990.

Employee\_of



# Relations

Person

Location

Organisation

Year

Mark lives in London. He has worked at University of London since 1990.



Based\_in

# Examples of IE systems

# Health and Safety Information Extraction (HaSIE)

The screenshot shows a software window titled "Hse" with a sidebar on the left containing a tree view with the following items: "CompanyName", "HSEParagraphs", "Awards", and "Accidents".

- CompanyName:** A text box containing "BAA".
- HSEParagraphs:** A large text area containing a paragraph: "sustainability management system. ... BAA has received a RoSPA gold award for occupational safety for the fourth year running. The award is given only if a consistently good or continuously improving performance can be demonstrated over a four-year period. The accident frequency ratio for construction projects was 0.4 (0.49) per 100,000 hours worked, less than one third of the national accident frequency rate in the construction sector. The company is running a ?One in a Million? campaign to raise safety consciousness and standards in construction and reduce the accident frequency rate still further to one for every million man hours worked. ... We have no higher priority than the safety and security of the passengers, staff and organisations that use our airports. In order to ensure that our systems and practices are continually assessed and upgraded, we work".
- Awards:** A text box containing "BAA has received a RoSPA gold award".
- Accidents:** A text box containing "The accident frequency ratio for construction projects was 0.4 (0.49) per 100,000 hours worked, less than one third of the national accident frequency rate in the construction sector."

At the bottom of the window, there is a status bar with the text "Record: " followed by navigation icons (back, forward, search, etc.) and the text "1 of 36".

# Obstetrics records

The screenshot displays the GATE software interface with the following components:

- Menu Bar:** File, Options, Tools, Help.
- Toolbar:** Icons for file operations, pipeline management, and search.
- Left Panel (GATE):**
  - Applications
    - pipeline
    - Language Resources
  - Processing Resources
    - Cleanup
    - Annotation Set Trajectory
    - IE Transducer
    - Flexible Gazetteer
    - Roots gazetteer
- Annotations List (Bottom Left):**

Property	Value
MimeType	text/
currentGravidity	3
day	20
gate.SourceURL	file:/
month	8
shift	12
- Main Text Area:**

1:30pm

Cx: 3cm contractions q2-3min. FHR: reassuring. reactive.

4:00pm

BP: 140/90

PV: 6cm, 60%, -1; soft consistency; anterior position; cephalic; Intact membranes; no vaginal bleeding.

Contractions: 3/10min; regular; moderate

On urinalysis: Protein > 300mg

BP before 20 weeks gestation: 120/80

Plan: monitor Vital Signs by protocol for elevated BP

5:15pm
- Right Panel (Annotations):**

Annotation	Status
CesareanSectionInPriorDelivery	<input type="checkbox"/>
DiastolicBloodPressure	<input checked="" type="checkbox"/>
DiastolicBloodPressureBefore20W	<input checked="" type="checkbox"/>
Dinoprostone	<input checked="" type="checkbox"/>
EstimatedFetalWeight	<input type="checkbox"/>
FHREvaluation	<input type="checkbox"/>
GBSNeonatalSepsisAfterAPrevious	<input type="checkbox"/>
Gravidity	<input type="checkbox"/>
HighRiskForAnaphylaxis	<input type="checkbox"/>
MagnesiumSulfate	<input type="checkbox"/>
MembranesStatus	<input checked="" type="checkbox"/>
MyastheniaGravis	<input type="checkbox"/>
PatientAge	<input type="checkbox"/>
PelvicAdequacy	<input type="checkbox"/>
PenicillinAllergy	<input type="checkbox"/>
PreviousCesareanSectionType	<input type="checkbox"/>
SystolicBloodPressure	<input checked="" type="checkbox"/>
SystolicBloodPressureBefore20We	<input checked="" type="checkbox"/>
TimeStamp	<input type="checkbox"/>
UrineProtein	<input checked="" type="checkbox"/>
- Bottom Panel:**
  - Document Editor
  - Initialisation Parameters

# Guided tour to GATE GUI

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

# ANNIE tool

You can try ANNIE tool online at:

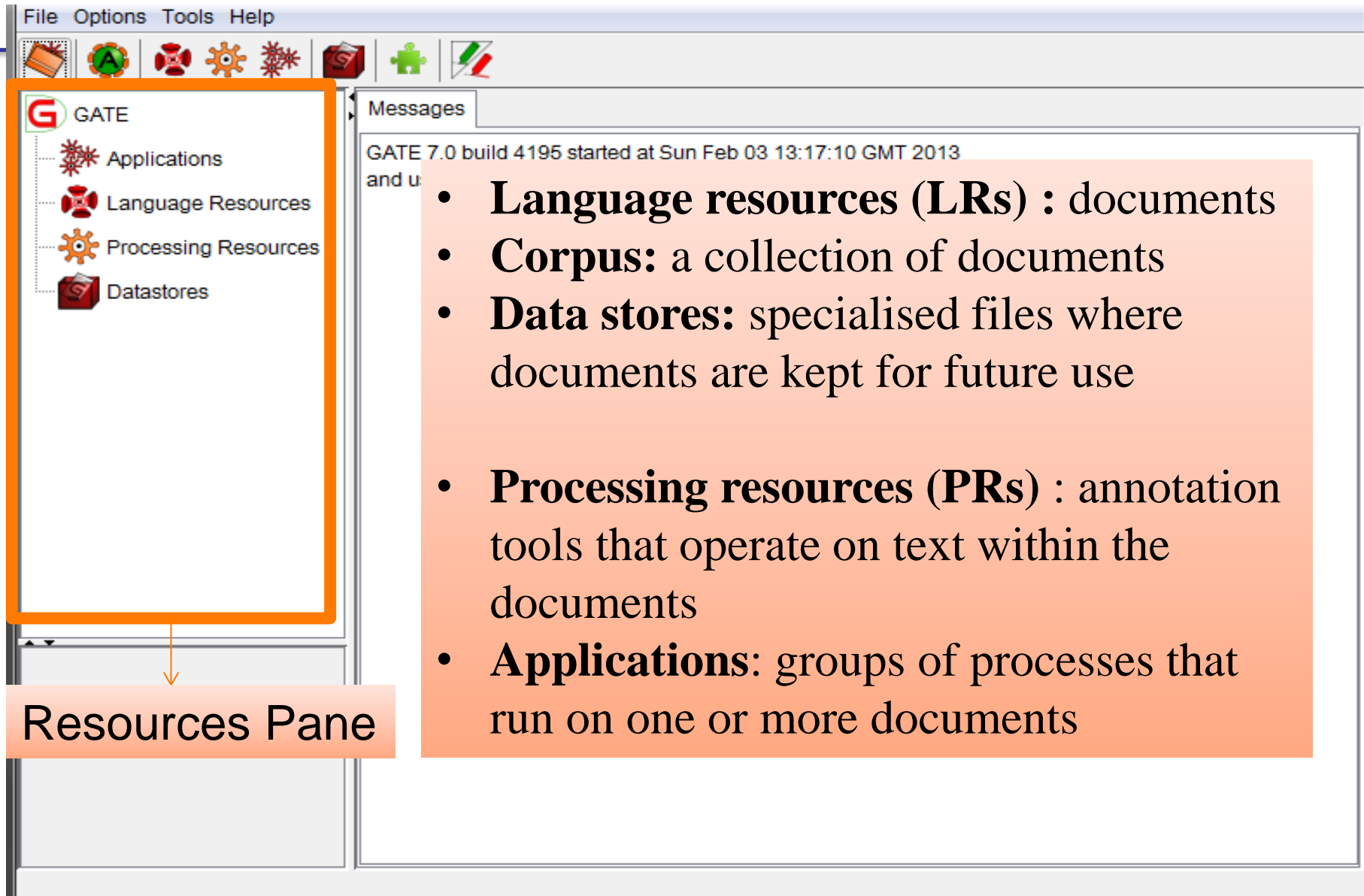
- <http://services.gate.ac.uk/annie/>

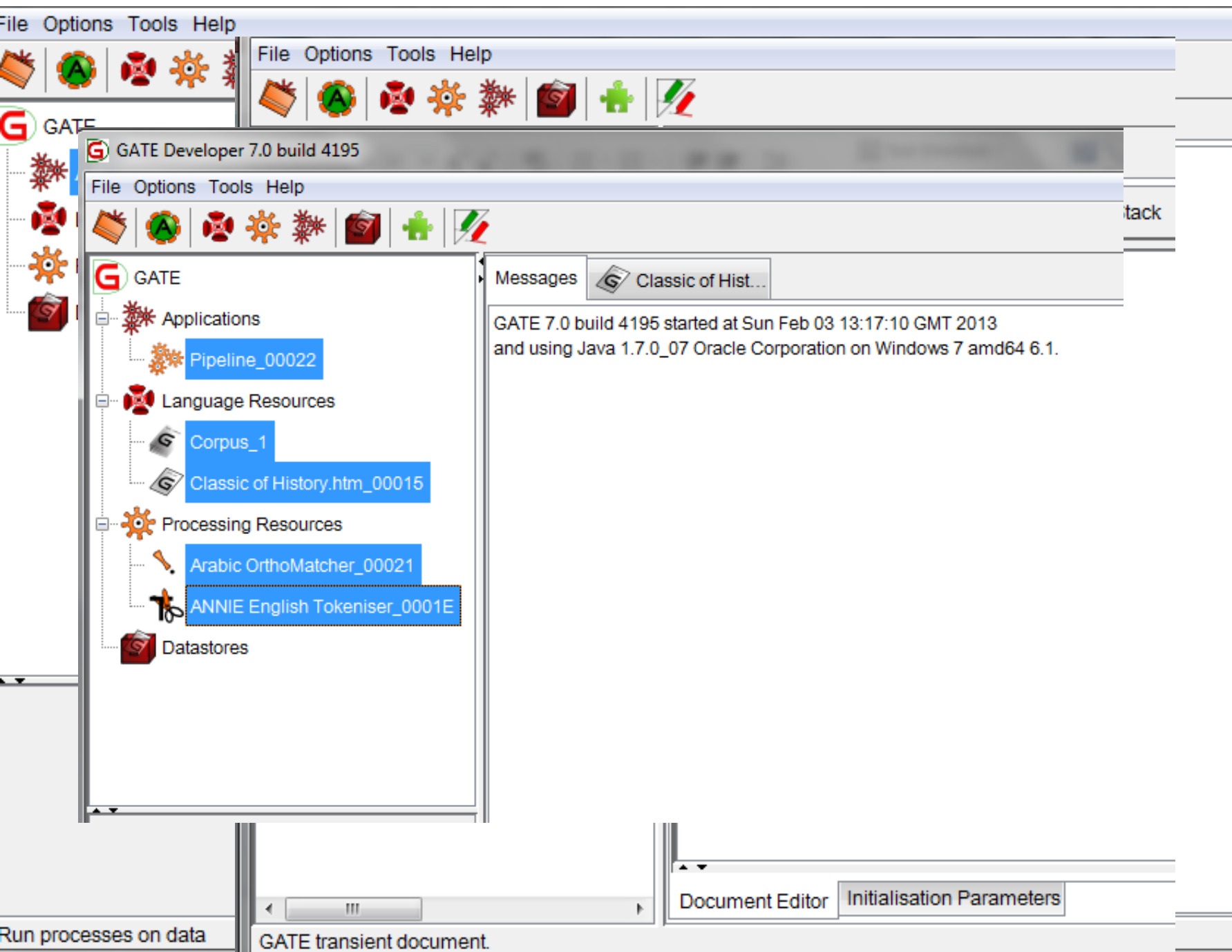


## How we will work on tutorial

- Each section in this tutorial will be explained first. After that enough time will be given to you to try yourself.
- red texts indicates your time to try experimenting with GATE.

# Guided tour to GATE GUI

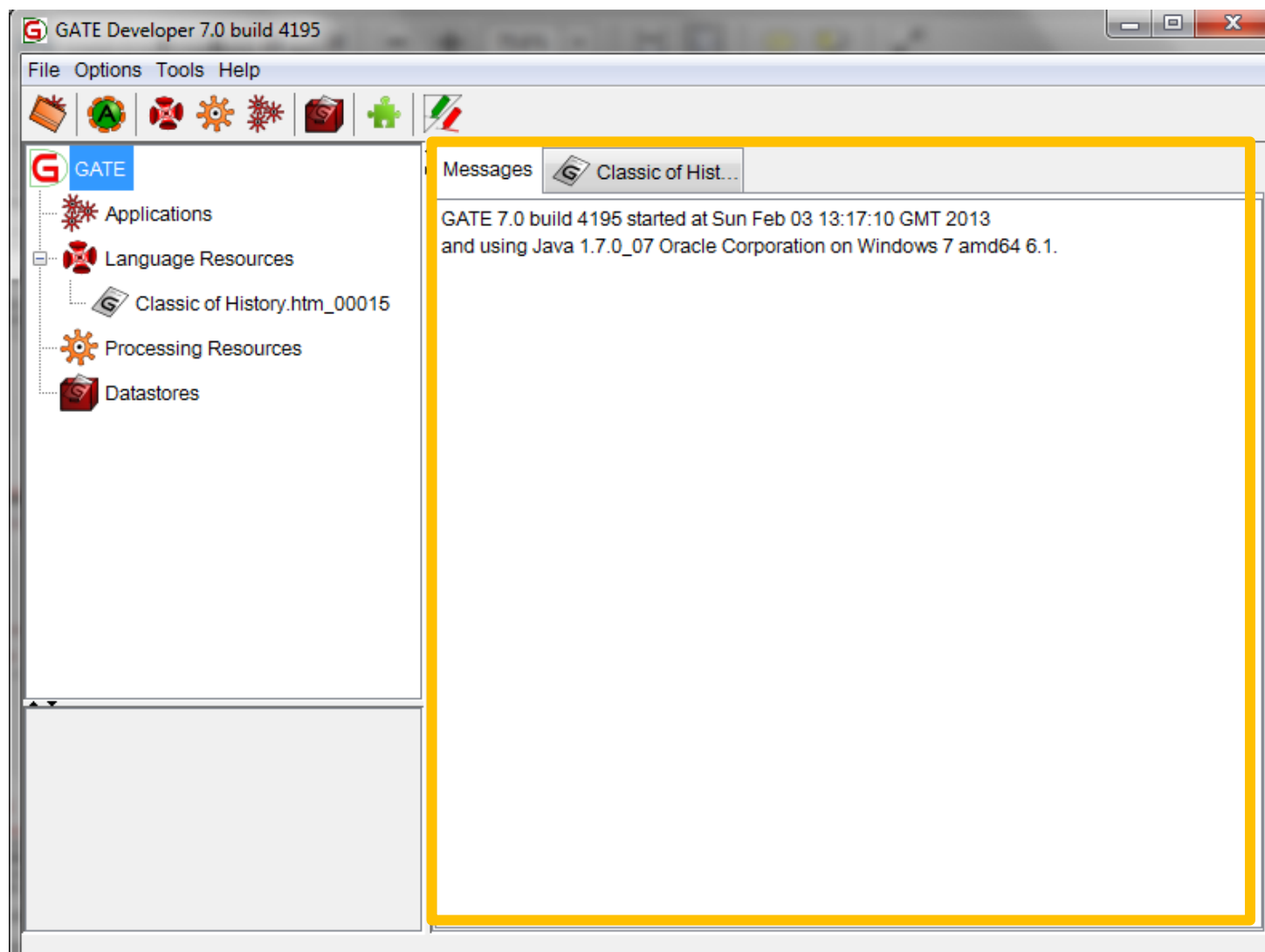




# Parameters

- Applications, LR's, and PR's all have various parameters.
- Parameters enable different settings to be used, e.g. case sensitivity
- **Initialisation Parameters** (set at load time): cannot be changed without reloading.
- **Run time Parameters:** can be changed between each application run

# Display Pane



GATE Developer 7.0 build 4195

File Options Tools Help

GATE

- Applications
- Language Resources
  - Classic of History.htm\_00015
- Processing Resources
- Datastores

Example of LRs

Messages Classic of Hist...

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

The Classic of History was one of the five texts on which Confucian teaching was based. The classic of history recorded the deeds of the sage kings.

Chronologies

China: Early China

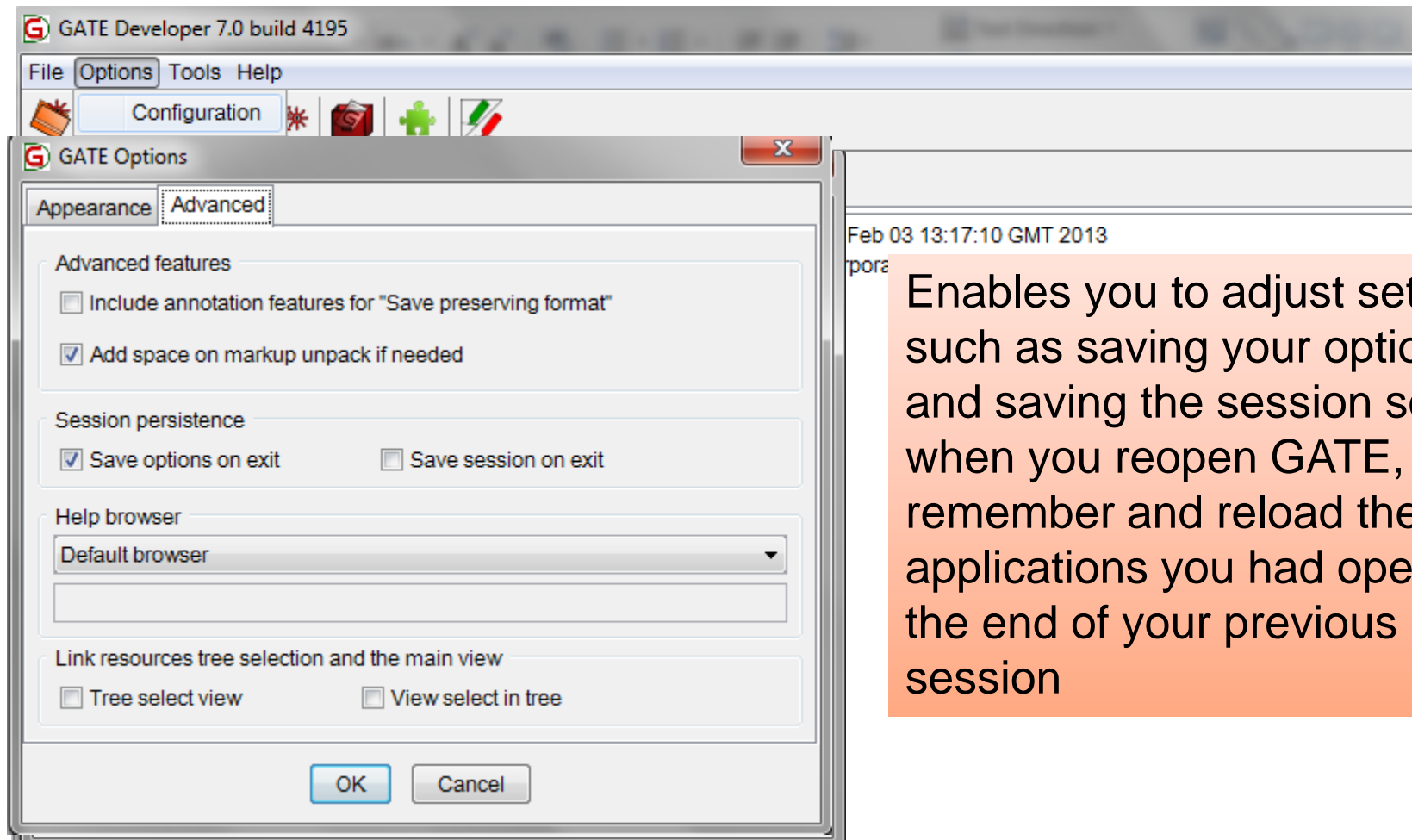
In the twelfth month of the first year . . . Yi Yin sacrificed to the former king, and presented the heir-king reverently before the shrine of his grandfather. All the princes from the domain of the nobles and the royal domain were present; all the officers also, each continuing to discharge his particular duties, were there to receive the orders of the chief minister. Yi Yin then clearly described the complete virtue of the Meritorious Ancestor for the instruction of the young king.

He said, "Oh! of old the former kings of Xia cultivated earnestly their virtue, and then there were no calamities from Heaven. The spirits of the hills and rivers alike were all in tranquility; and the birds and beasts, the fishes and tortoises, all enjoyed their existence according to their nature. But their descendant did not follow their example, and great Heaven sent down calamities, employing the agency of our ruler—who was in

Document Editor Initialisation Parameters

C	MimeType	▼	text/html
C	gate.SourceURL	▼	file:/C:/Users/2012
C		▼	

# Setting up GATE options



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

## Try out GATE

- Open GATE

Start → All Programs → GATE developer 7.0

- Try setting different options in GATE

Click Options → Configuration → Appearance

- Download the presentation file.
- Download the “Hands-on-materials.zip” file from

<https://sites.google.com/site/nlelab2013/gate>

Save the zipped file on desktop and unzip it.

The folder “Hands-on-materials” contains all files required for the next experiments.



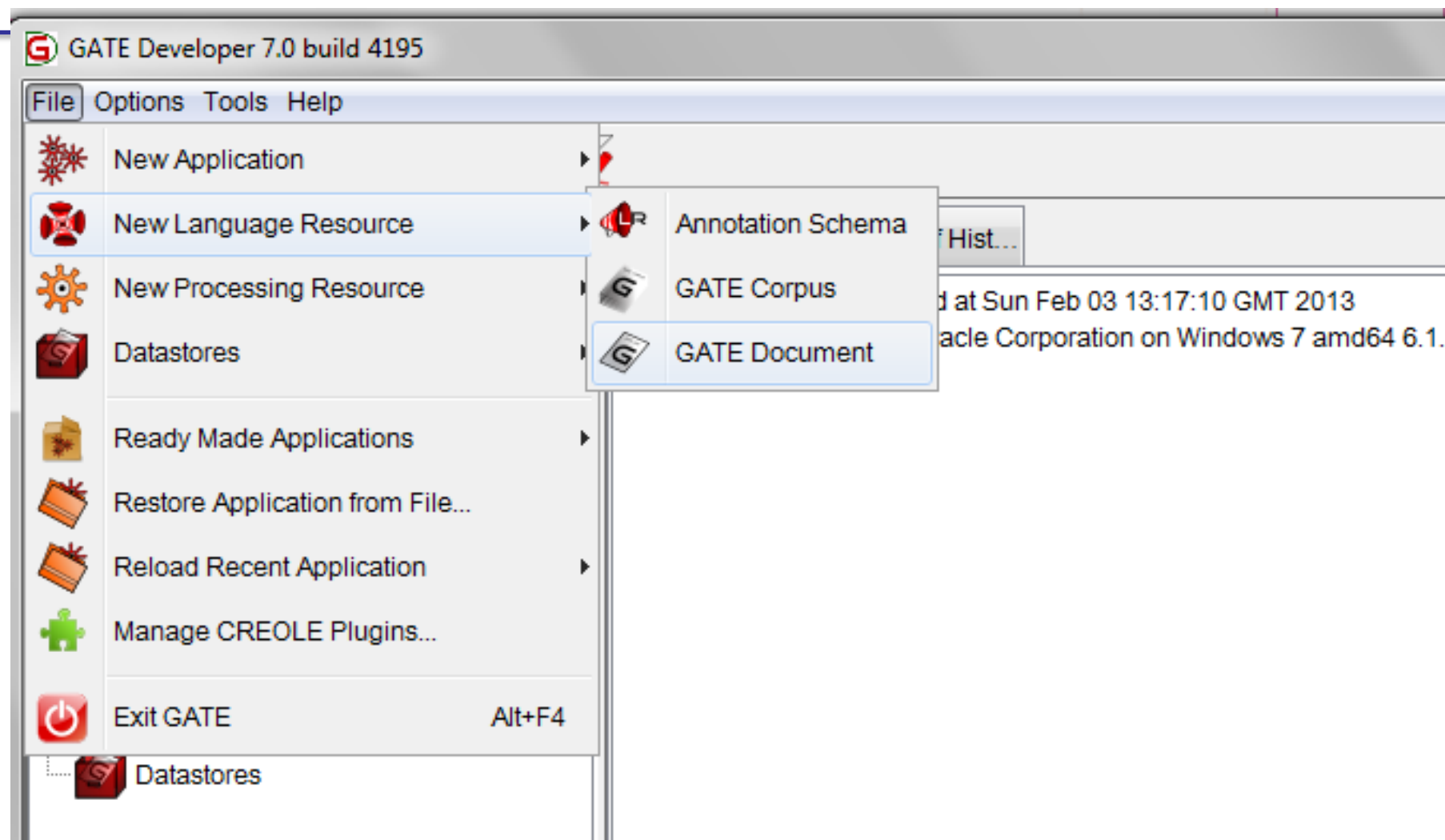
# Loading and Viewing Documents

- Loading a document and setting its parameters.
- Creating a corpus .
- Populating a corpus of documents in different ways
- Removing documents.

# Loading and Viewing Documents

- GATE can process documents in all kinds of formats: plain text, HTML, XML, PDF, Word etc.
- When GATE loads a document, it converts it into a special format for processing.
- Documents can be exported in various formats or saved in a datastore for future processing within GATE.

# Loading Document



# Document Initialisation parameters

Parameters for the new GATE Document

Name:

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

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.

# Document Initialisation parameters

Parameters for the new GATE Document

Name:

Name	Type	Required	Value
collectRepositioningInfo	Boolean	✓	false
encoding	String		
markupAware	Boolean	✓	true
mimeType	String		
preserveOriginalContent	Boolean	✓	false
<b>sourceUrl</b>	<b>URL</b>	✓	<input type="text"/>
sourceUri	Long		
<b>stringContent</b>	Long		
sourceOnStartOnSet	Long		

OK Cancel Help

You can also just type a string of text into the box by selecting **stringContent** rather than **sourceUrl**.

# Document Initialisation parameters

Parameters for the new GATE Document

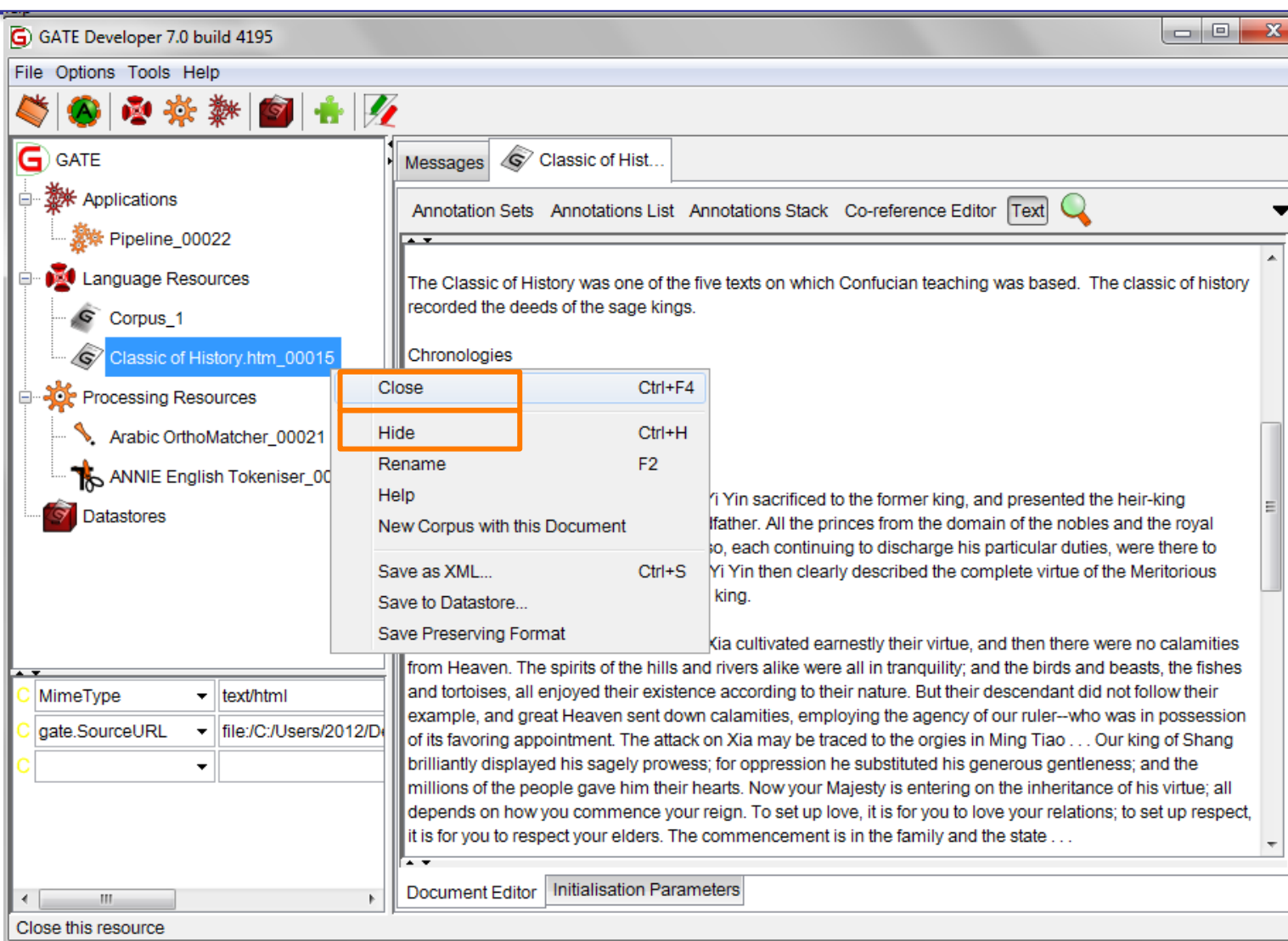
Name:

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

OK Cancel Help

Set to true to ensure GATE will process any existing annotations such as HTML tags and present them as annotations rather than leaving them in the text.

# Opening and closing the document



# Viewing the document

GATE Developer 7.0 build 4195

File Options Tools Help

GATE

- Applications
  - Pipeline\_00022
- Language Resources
  - Corpus\_1
    - Classic of History.htm\_00015
- Processing Resources
  - Arabic OrthoMatcher\_00021
  - ANNIE English Tokeniser\_0001E
- Datastores

Messages Classic of Hist...

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Toggle the view of Annotation Sets F3

The Classic of History

The Classic of History was one of the five texts on which Confucian teaching was based. The classic of history recorded the deeds of the sage kings.

Chronologies

China: Early China

Type	Set	Start	End	Id	Features
h1	Original markups	81	103	49	{}

1 Annotations (0 selected) Select:

Document Editor Initialisation Parameters

body div font h1 head hr html i img link meta p span strong style table tbody td title tr

New

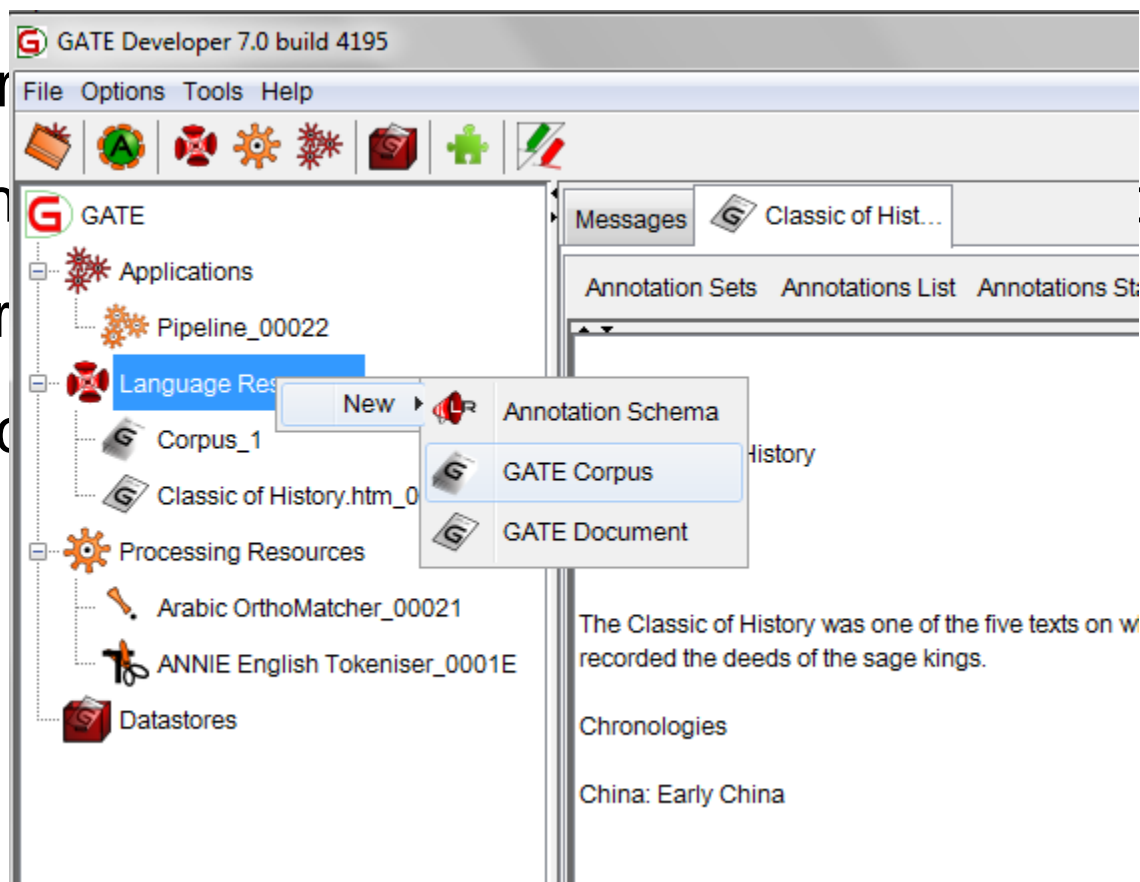


## Experiment 2 – Opening the document in GATE

- load the document “Anna Comnena Alexiad.htm” from “hands-on-materials” folder.
  - right click on Language Resources and select “New → GATE Document” or
  - File menu → New Language Resource → GATE Document
- A dialogue box will appear.
- Leave the name input box empty.
- Click the file browser icon to navigate to the correct document.
- To view a document, double click on the document name in the Resources pane
- To view the annotations, you first need click “Annotation Sets”, and then select the relevant set and annotation(s) on the right hand side of the GUI
- To see a list of annotations at the bottom, click on “Annotations List”

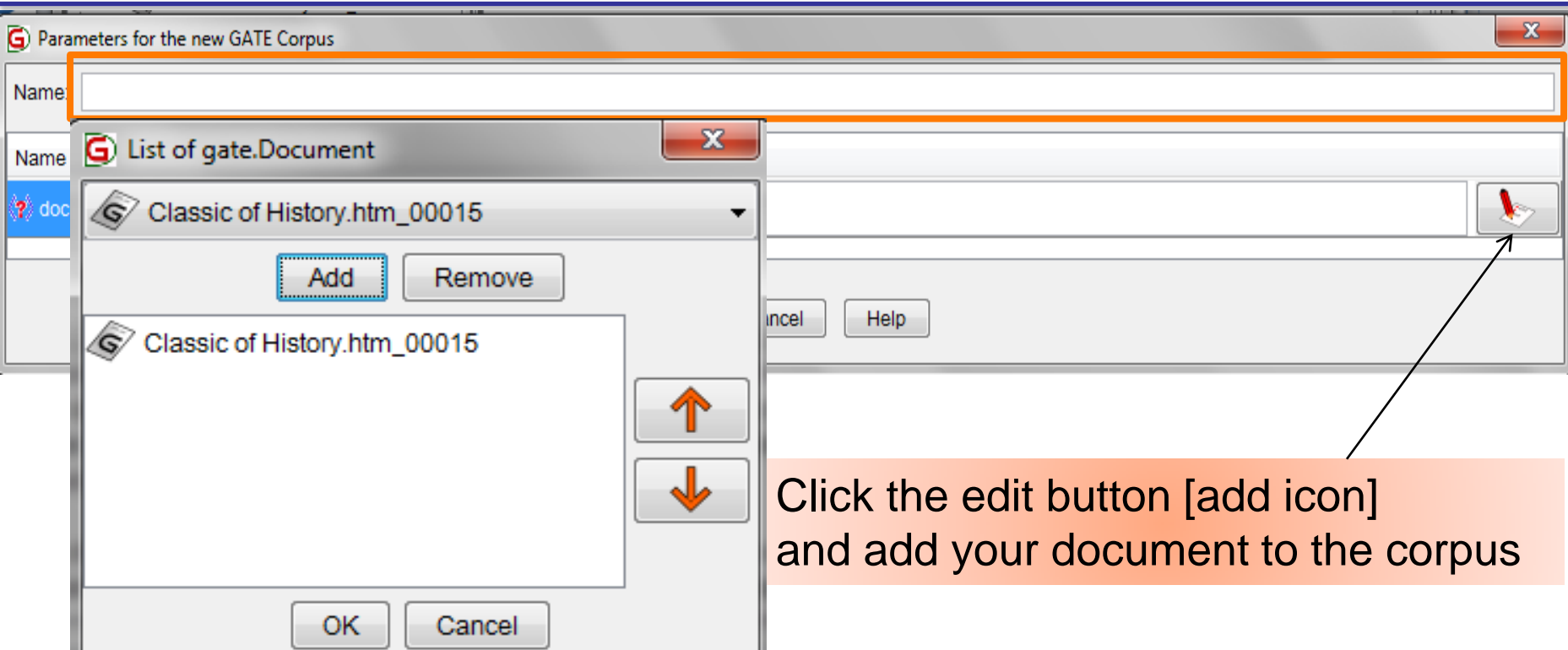
# Creating a corpus

- A corpus is a collection of texts
- For many NLP tasks, you need to work with a corpus that contains many documents



work with  
even if

# Corpus Initialisation Parameters



# Another way to add documents to a corpus

File Options Tools Help

GATE

- Applications
  - Pipeline\_00022
- Language Resources
  - Anna Comnena Alexiad.htm\_00028
  - Germanic Law.htm\_00027
  - Corpus\_1**
  - Classic of History.htm\_00015
- Processing Resources
  - Datastores

**((1))**  
Double click on the corpus

Messages Classic of Hist... Corpus\_1

**((2))**  
Press on add icon

Add new document(s) to this corpus

- use the toolbar buttons at the top of this view
- drag documents from the left resources tree and drop them below
- right click on the corpus in the resources tree and choose 'Populate'

Index Document name

Add document(s) to this corpus

- Anna Comnena Alexiad.htm\_00028
- Germanic Law.htm\_00027
- Classic of History.htm\_00015**

OK Cancel

Corpus editor Initialisation Parameters Corpus Quality Assurance

# Removing documents from a corpus

File Options Tools Help

G GATE

- Applications
  - Pipeline\_00022
- Language Resources
  - Anna Comnena Alexiad.htm\_00028
  - Germanic Law.htm\_00027
  - Corpus\_1**
  - Classic of History.htm\_00015
- Datastores

Messages Classic of Hist... Corpus\_1

All the documents loaded in the system

(1) Double click on the corpus

Index	Document name
0	Classic of History.htm_00015
1	Anna Comnena Alexiad.htm_00028
2	Germanic Law.htm_00027

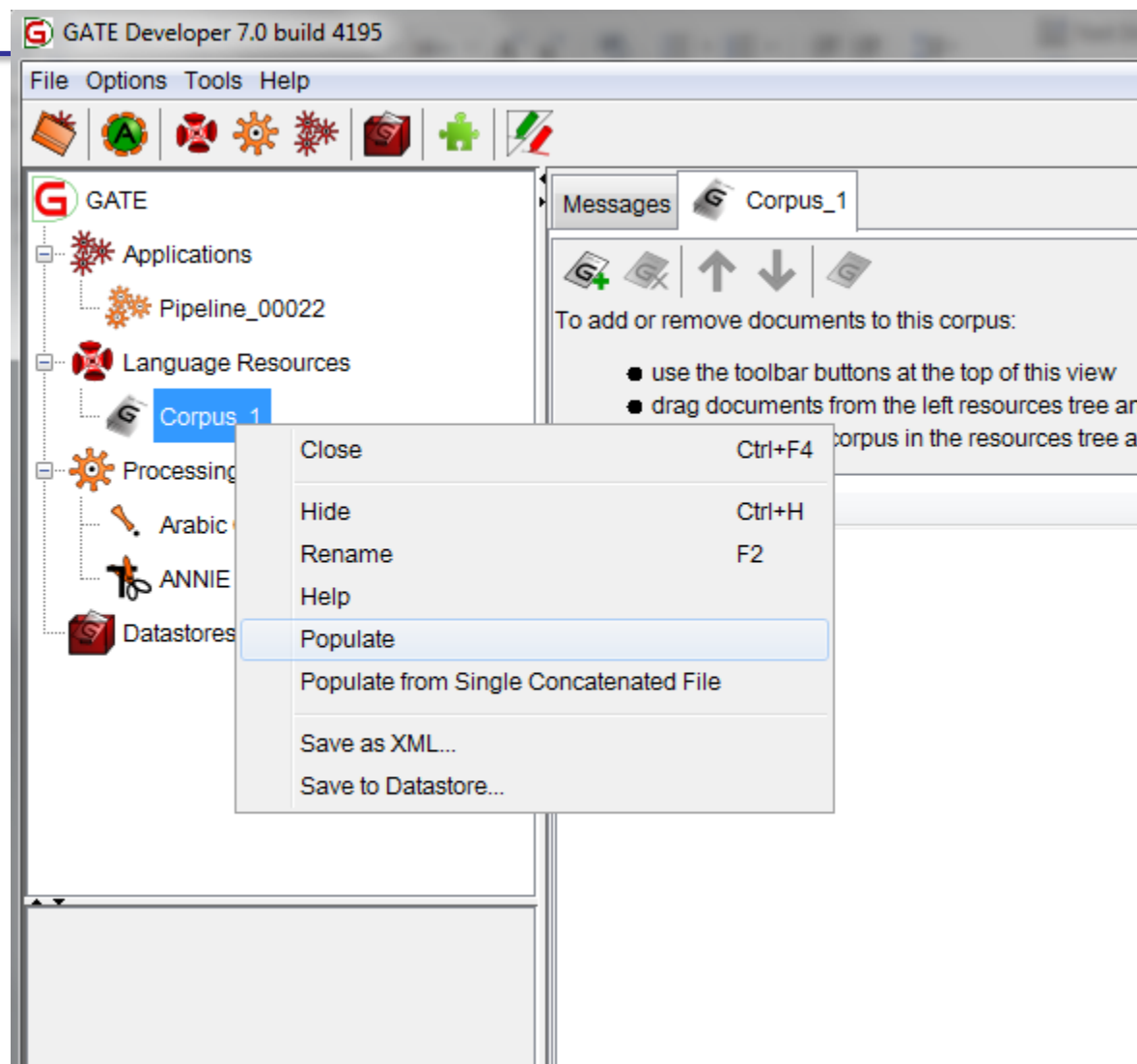
(2) Select the document you want to delete

(3) Press on delete icon

## Removing documents from a corpus

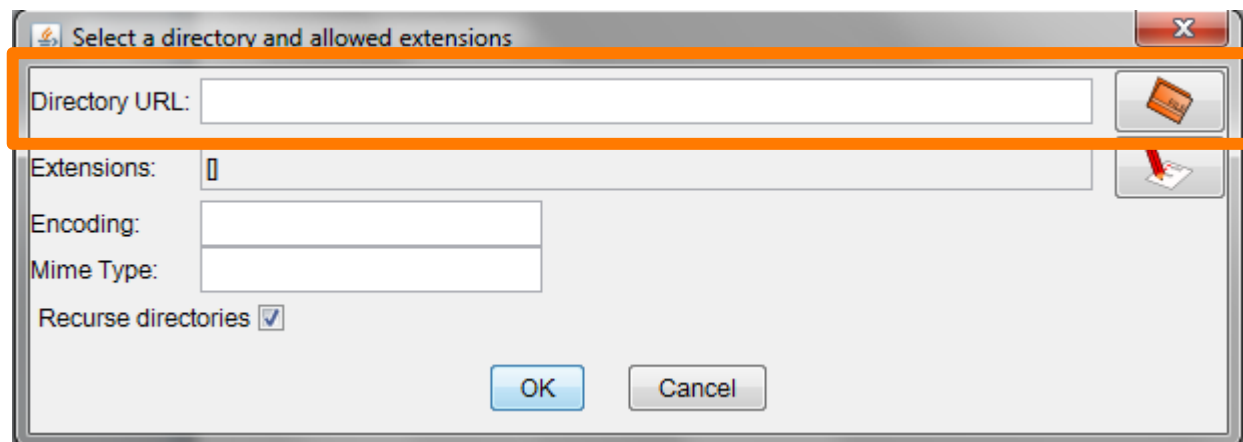
- To remove documents from a corpus, use the X button in the corpus editor
  - Note that this does not remove the document from GATE, just from the corpus
  - The document is available to be added to other corpora. Indeed a document can belong to several corpora
- If you do remove the document from GATE, it will also remove it from the corpus
- But if you remove the corpus, it doesn't remove the document!

# Populate the corpus in one go



you don't have to  
and allows you to  
in one go.

# Populate the corpus in one go



- **Extensions parameter:** lets you select only documents of a certain type e.g., “xml” or “htm” files.
- **Encoding:** lets you choose the right encoding for the documents. The wrong encoding can cause characters to be incorrectly displayed e.g., “UTF-8”
- **Recurse directories :** load documents in any subdirectories



## Experiment 3: creating a corpus and populate it

- Create a corpus
  - Right click Language Resources → New → GATE Corpus.
- A dialogue box will appear.
- Leave a name input box empty and press OK.
- Right click on the created corpus in the Resources pane and select Populate.
- A dialogue box will appear
- Use the file browser icon to select the name of the directory in which your documents are stored
- In Extensions parameter ,type “htm” in the box (without the quotes)
- In Encoding parameter , enter “UTF-8” .
- Press OK
- all the documents will be loaded in one go
- Double click the corpus to view it.

# Preprocessing Resources

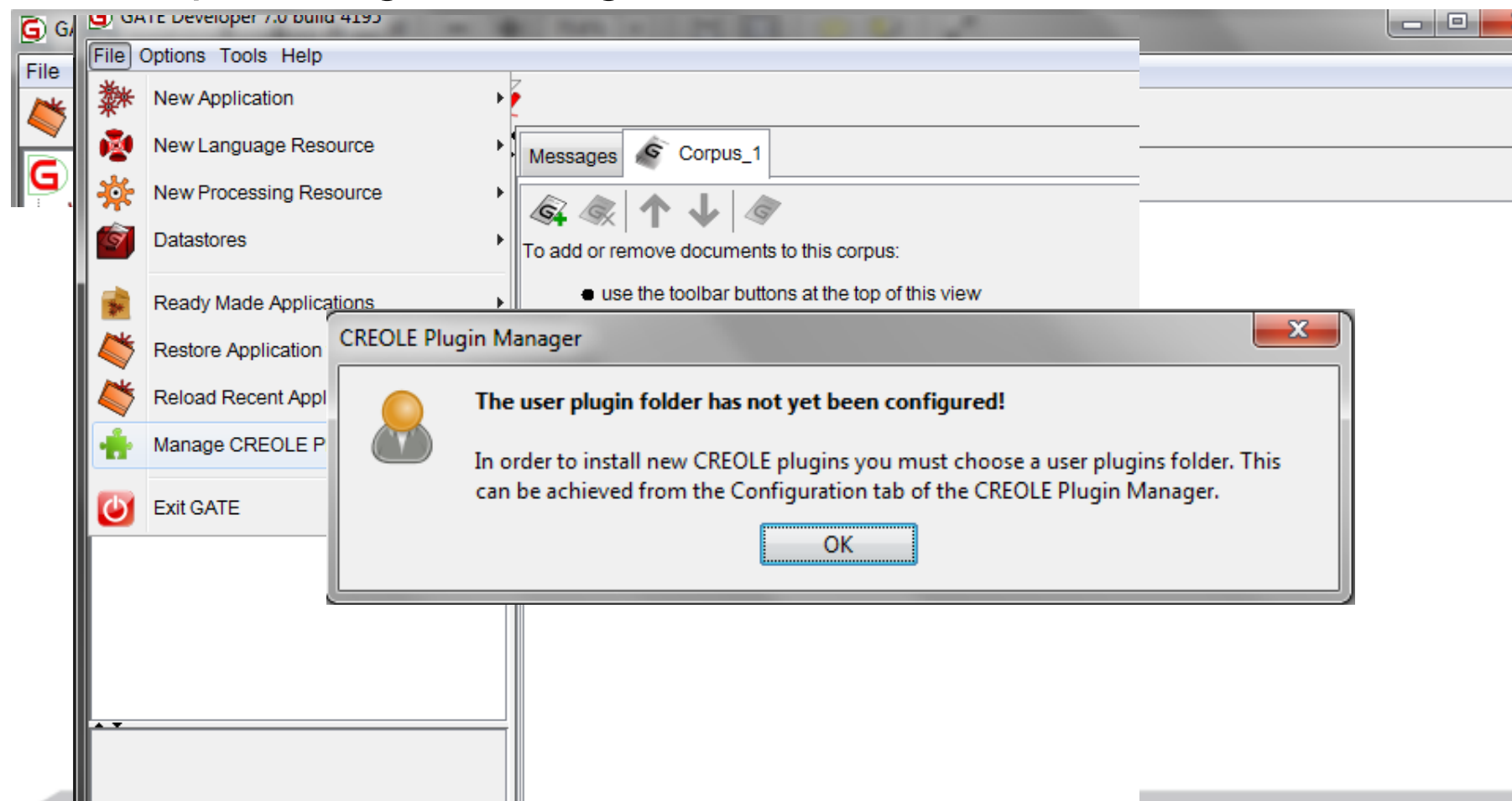
- Loading processing resources.
- Managing plugins.
- Loading and running ANNIE and pre-existing applications
- Creating a new application

# Preprocessing Resources

- Processing resources (PRs) are the tools that creates or modifies annotations on the text .They implement algorithms.
- 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.
- 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

To open Plugin Manager:



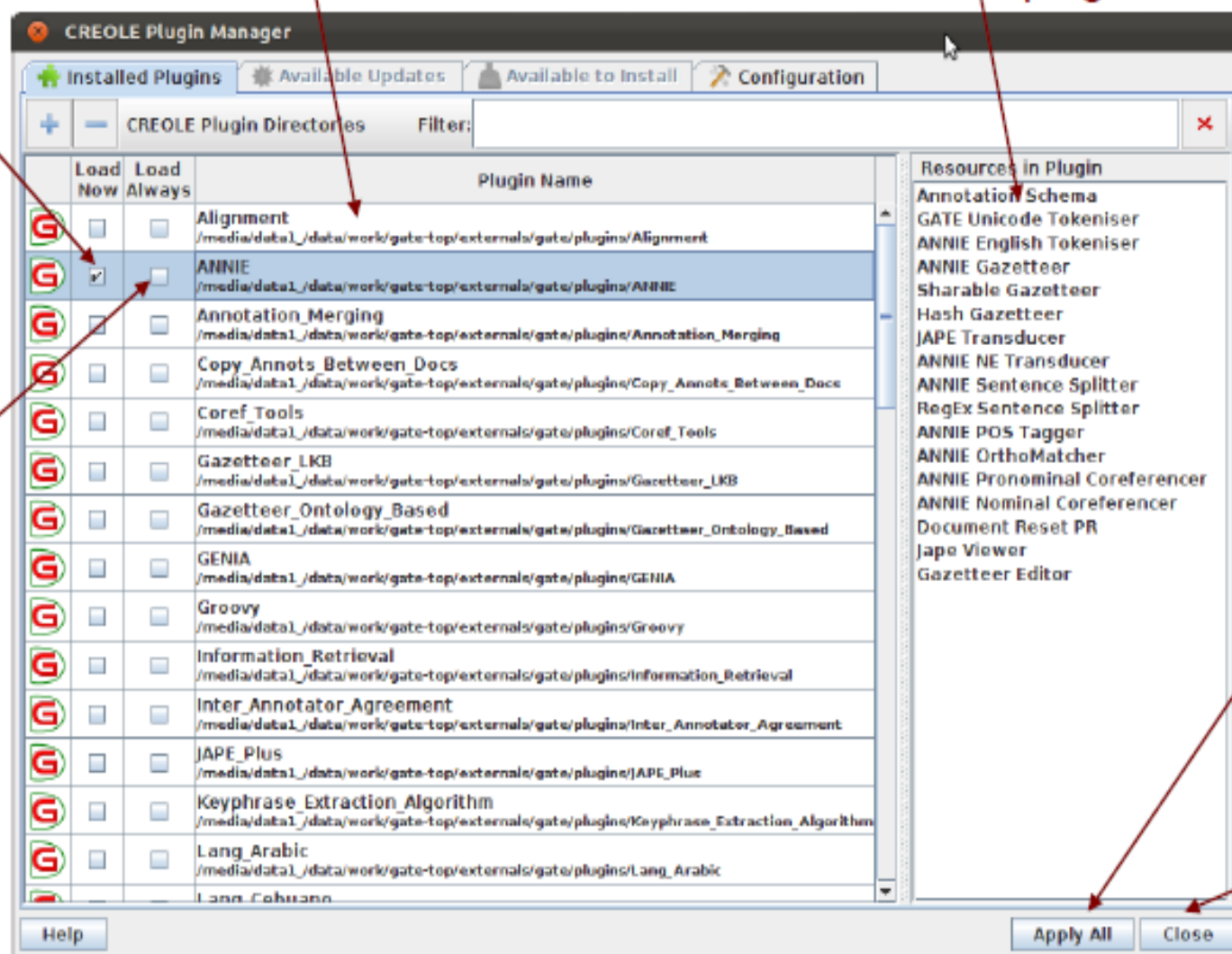
# Plugins

Load the plugin for this session only

Load the plugin everytime GATE starts

List of available plugins

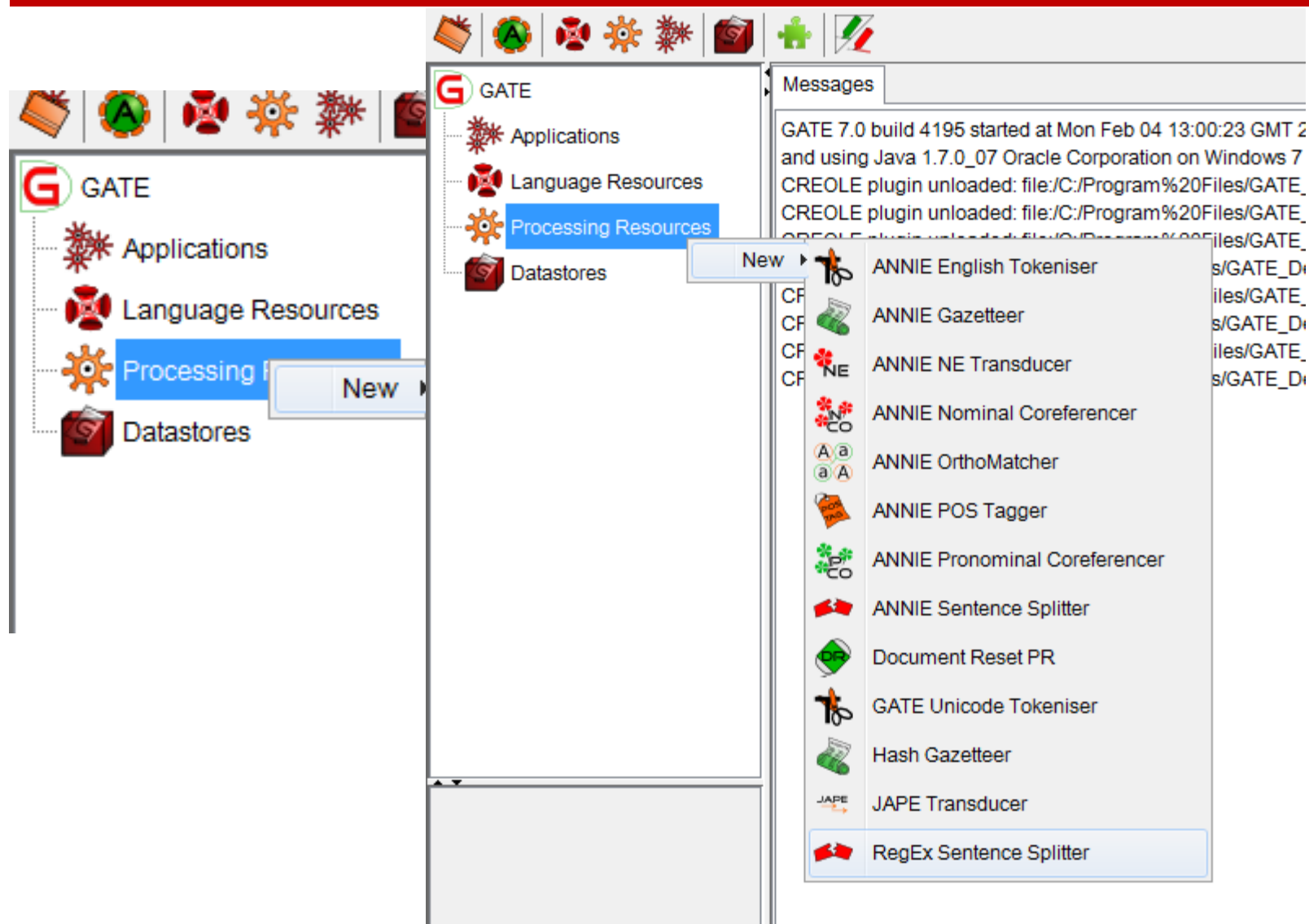
Resources in the selected plugin



Apply all the settings

Close the plugins manager

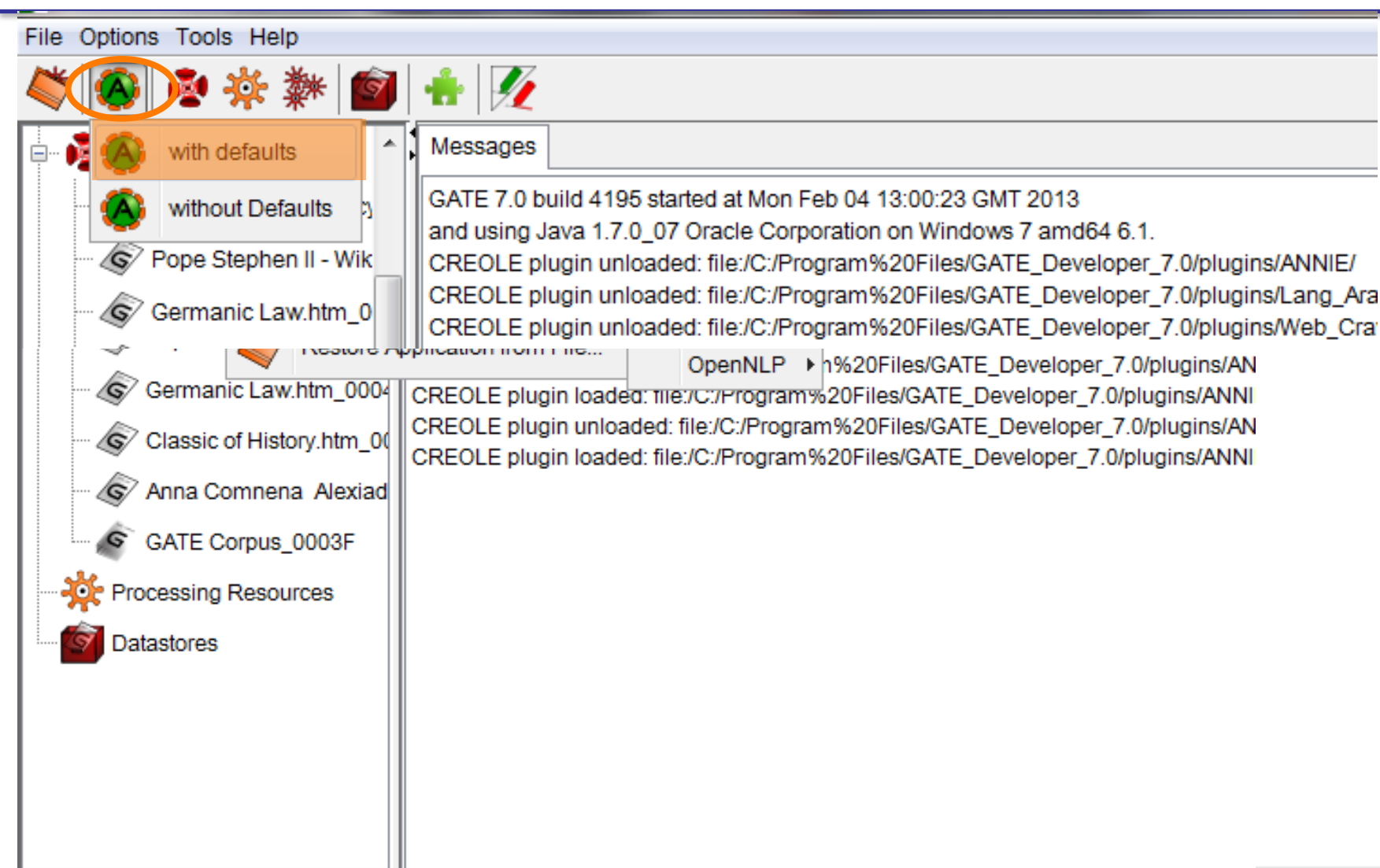
# Plugins



# Applications

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

# ANNIE Tool





# Running ANNIE Tool

- Double click on the ANNIE application to view it.

The screenshot shows the ANNIE Tool interface with several key components highlighted by orange boxes and arrows:

- Loaded Processing resources:** An empty table with columns 'Name' and 'Type'.
- Selected Processing resources:** A table listing the execution pipeline. The 'ANNIE POS Tagger' is selected and highlighted in blue.
 

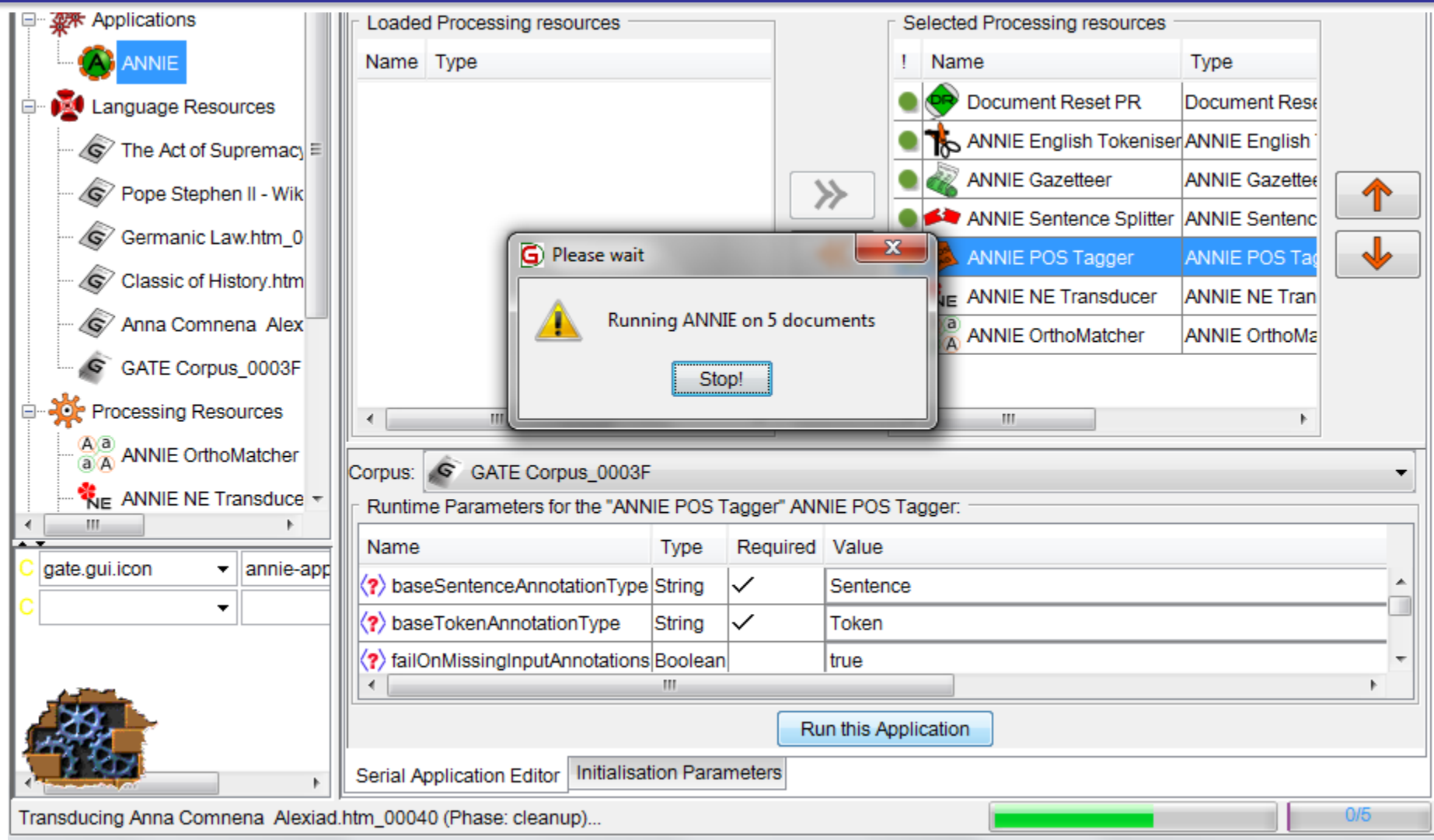
! Name	Type
Document Reset PR	Document Reset
ANNIE English Tokeniser	ANNIE English
ANNIE Gazetteer	ANNIE Gazetteer
ANNIE Sentence Splitter	ANNIE Sentence
<b>ANNIE POS Tagger</b>	<b>ANNIE POS Tag</b>
ANNIE NE Transducer	ANNIE NE Tran
ANNIE OrthoMatcher	ANNIE OrthoMa
- Corpus:** A dropdown menu showing 'GATE Corpus\_0003F'.
- Runtime Parameters for the 'ANNIE POS Tagger' ANNIE POS Tagger:** A table showing configuration options.
 

Name	Type	Required	Value
baseSentenceAnnotationType	String	✓	Sentence
baseTokenAnnotationType	String	✓	Token
failOnMissingInputAnnotations	Boolean		true
- Run this Application:** A button at the bottom center of the interface.

Annotations on the right side of the interface:

- PRs in order of their execution (points to the Selected Processing resources table)
- Corpus on which the application is executed (points to the Corpus dropdown)
- Runtime parameters of the selected PR (points to the Runtime Parameters table)
- Execute the application (points to the Run this Application button)

# ANNIE Tool



The screenshot shows the ANNIE Tool interface. A modal dialog box titled "Please wait" is displayed in the center, containing a yellow warning icon and the text "Running ANNIE on 5 documents" with a "Stop!" button. The background interface includes a left sidebar with a tree view of resources, a main area with "Loaded Processing resources" and "Selected Processing resources" tables, and a bottom section for runtime parameters and a progress bar.

**Left Sidebar:**

- Applications: ANNIE
- Language Resources:
  - The Act of Supremacy
  - Pope Stephen II - Wik
  - Germanic Law.htm\_0
  - Classic of History.htm
  - Anna Comnena Alex
  - GATE Corpus\_0003F
- Processing Resources:
  - ANNIE OrthoMatcher
  - ANNIE NE Transducer

**Main Area:**

**Loaded Processing resources:**

Name	Type
------	------

**Selected Processing resources:**

!	Name	Type
●	Document Reset PR	Document Rese
●	ANNIE English Tokeniser	ANNIE English
●	ANNIE Gazetteer	ANNIE Gazettee
●	ANNIE Sentence Splitter	ANNIE Sentenc
●	ANNIE POS Tagger	ANNIE POS Tag
●	ANNIE NE Transducer	ANNIE NE Tran
●	ANNIE OrthoMatcher	ANNIE OrthoMa

**Bottom Section:**

Corpus: GATE Corpus\_0003F

Runtime Parameters for the "ANNIE POS Tagger" ANNIE POS Tagger:

Name	Type	Required	Value
baseSentenceAnnotationType	String	✓	Sentence
baseTokenAnnotationType	String	✓	Token
failOnMissingInputAnnotations	Boolean		true

Run this Application

Serial Application Editor Initialisation Parameters

Transducing Anna Comnena Alexiad.htm\_00040 (Phase: cleanup)...

0/5

# Viewing the results

The screenshot shows the GATE software interface. On the left, a sidebar lists project resources including 'Applications', 'Language Resources', and 'Processing Resources'. The main window displays a text document titled 'Pope Stephen II...' with several words highlighted in green: 'Stephen', 'Ravenna', and 'Pepin'. An orange callout box with the text '((3)) click on any Annotation types in the Default (unnamed) set' points to the 'Annotations List' tab. Below the text, a table lists the annotations:

Type	Set	Start	End	Id	Features
Person		5	27	9356	{gender=male, matches=[9356, 9357, 936
Person		57	67	9357	{gender=male, matchedWithLonger=true, i
Person		73	82	10340	{NMRule=Unknown, kind=PN, matches=[9
Person		196	207	9359	{gender=male, matches=[9359, 9360, 945
Person		223	230	9360	{gender=male, matchedWithLonger=true, i

Below the table, it says '283 Annotations (0 selected) Select:'. On the right, a sidebar lists various annotation types with checkboxes: Address, Date, FirstPerson, JobTitle, Location, Lookup, Organization, Person (checked), Sentence, SpaceToken, Split, Temp, Title, Token, Unknown, and UriPre. At the bottom right, there is a 'New' button.

# Viewing the results

Messages GATE Corpus\_000... ANNIE Pope Stephen II...

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

between the Byzantine Papacy and the Frankish Papacy .

Contents

- 1 Allegiance to Constantinople
- 2 Alliance with the Franks
- 3 Duchy of Rome and the Papal States
- 4 References
- 5 External links

[edit] Allegiance to Constantinople

The Lombards to the north of Rome had captured Ravenna , capital of the Eastern Roman Empire Exarchate of Ravenna , in 751, and began to put pressure on the city of Rome.

Relations were very strained in the mid-8th century between the papacy and the Eastern Roman emperors over the support of the Isaurian Dynasty for iconoclasm . Likewise, maintaining political control over Rome became untenable as the Eastern Roman Empire itself

Type	Set	Start	End	Id	Features
Token	ANNIEresult	1080	1085	2384	{category=VBD, kind=word, length=5, orth=lowercase, s
Token	ANNIEresult	1086	1088	2386	{category=TO, kind=word, length=2, orth=lowercase, s
Token	ANNIEresult	1089	1092	2388	{category=VB, kind=word, length=3, orth=lowercase, s
Token	ANNIEresult	1093	1101	2390	{category=NN, kind=word, length=8, orth=lowercase, s

2919 Annotations (1 selected) Select:

Document Editor Initialisation Parameters

Annotations List

- ☐ Address
- ☐ Date
- ☐ FirstPerson
- ☐ JobTitle
- ☐ Location
- ☐ Lookup
- ☐ Organization
- ☒ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Temp
- ☐ Title
- ☒ Token
- ☐ Unknown
- ☐ UriPre

## change the name of the annotation set

- 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

# change the name of the annotation set

## Loaded Processing resources

- 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.

## Selected Processing resources

!	Name	Type
	Document Reset PR	Document Reset P
	ANNIE English Tokeniser	ANNIE English Tok
	ANNIE Gazetteer	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence S
	ANNIE POS Tagger	ANNIE POS Tagge
	ANNIE NE Transducer	ANNIE NE Transdu
	ANNIE OrthoMatcher	ANNIE OrthoMatch

Corpus: GATE Corpus\_0003F

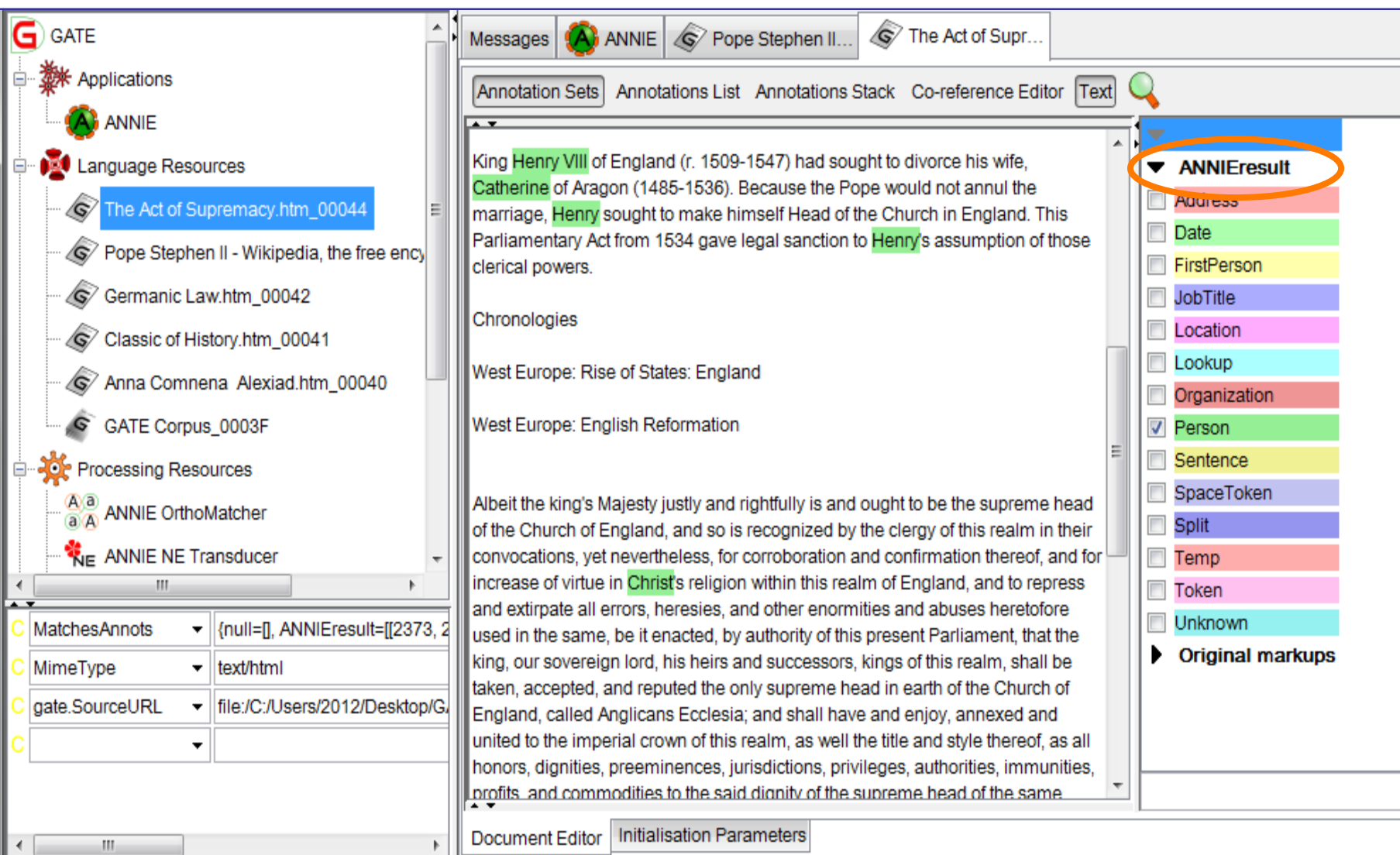
## Runtime Parameters for the "ANNIE Gazetteer" ANNIE Gazetteer:

Name	Type	Required	Value
annotationSetName	String		ANNIEresult
longestMatchOnly	Boolean	✓	true
wholeWordsOnly	Boolean	✓	true

Run this Application



# change the name of the annotation set



The screenshot shows the GATE software interface. On the left, the 'Language Resources' section is expanded, showing 'The Act of Supremacy.htm\_00044' selected. Below it, 'Processing Resources' includes 'ANNIE OrthoMatcher' and 'ANNIE NE Transducer'. The main window displays a text document with annotations. The 'Annotations List' tab is active, showing a list of annotations. The 'ANNIEresult' annotation set is highlighted with an orange circle. The 'Annotations List' tab is also highlighted with an orange circle. The 'Annotations List' tab shows a list of annotations, including 'ANNIEresult'.

Annotations List

Annotation Set	Annotation	Start	End	Text
ANNIEresult	Person	2373	2400	King Henry VIII
ANNIEresult	Person	2400	2430	Catherine
ANNIEresult	Person	2430	2460	Henry
ANNIEresult	Person	2460	2490	Henry

## Add a new Processing Resources

- Load a plugin called “Tools” using Plugins Manager.
- Right click on “Processing Resources” in the Resources Pane and select “New” → “ANNIE VP Chunker”
- Double click on ANNIE.
- You'll see the VP chunker in the list of loaded PRs. This means it's available in GATE, but isn't yet contained in the application.
- Add it to the application by selecting it and using the right arrow to transfer it.
- Now use the up arrow to move it to the right place in the application. It should go after (below) the POS tagger but before (above) the NE transducer.
- Change the inputASName and outputASName parameters to ANNIEresult.
- Run the application and view the results on the document.
- You should see a new annotation type “VG”.



# Saving documents

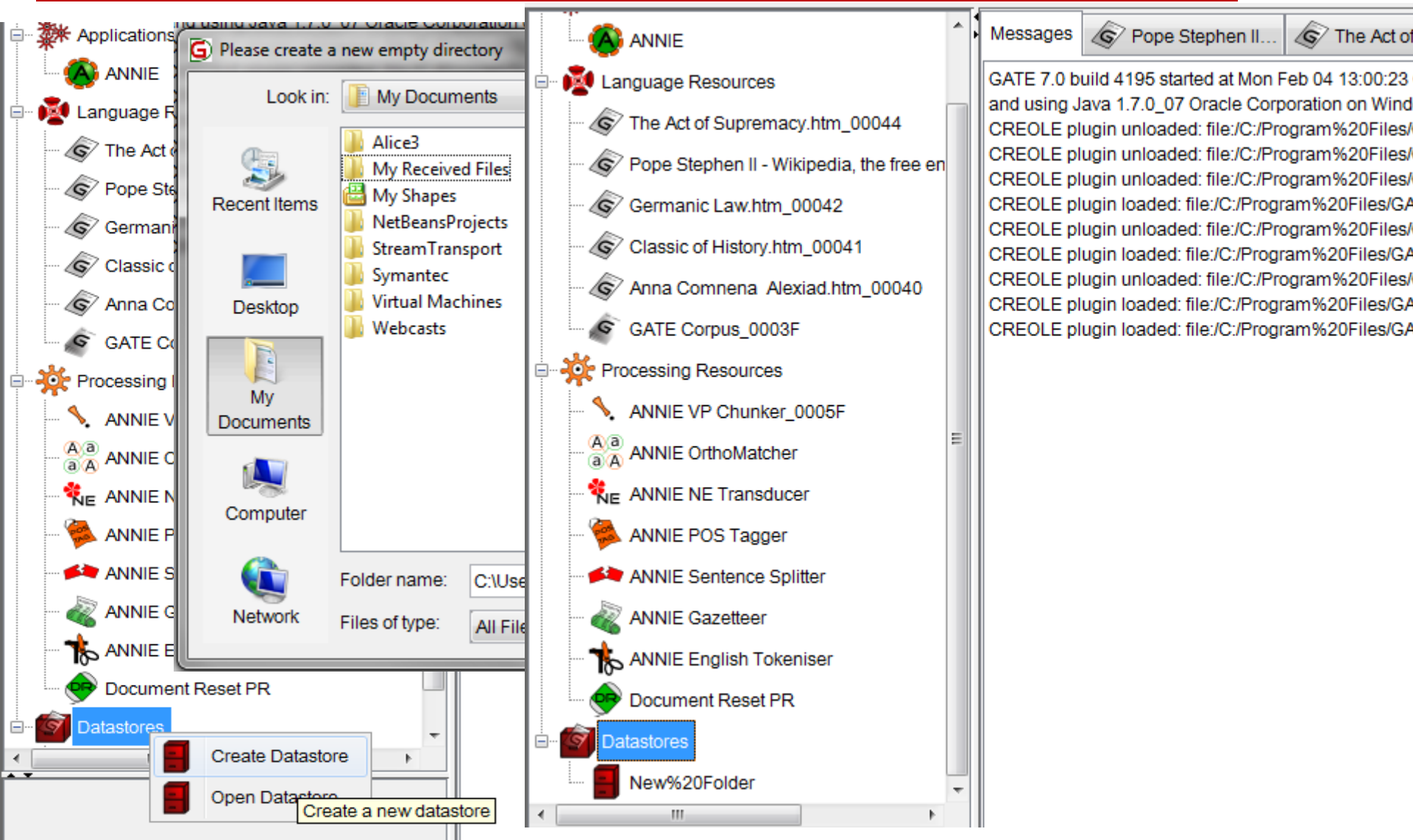
- Using datastores
- Saving documents for use outside GATE

# Saving documents

There are 2 types of datastore

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

# Create a new serial datastore



## 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

# remove things from the datastore

The screenshot shows the GATE software interface. On the left, the 'Datastores' section is expanded, showing a tree structure with 'New%20Folder' selected. An orange callout box labeled '(1)' points to this selection with the text: 'Double click on the name of the datastore'. On the right, the 'Messages' pane shows the contents of the selected folder, including 'GATE Document' and 'GATE Serial Corpus'. The 'GATE Serial Corpus' is expanded, showing a list of documents. The document 'GATE Corpus\_0003F' is selected, and an orange callout box labeled '(2)' points to it with the text: 'Double click on the name of corpus or documents and then select delete'. A context menu is visible over the selected document, with 'Delete' and 'Load' options.

**((1))**  
Double click on the name of the datastore

**((2))**  
Double click on the name of corpus or documents and then select delete

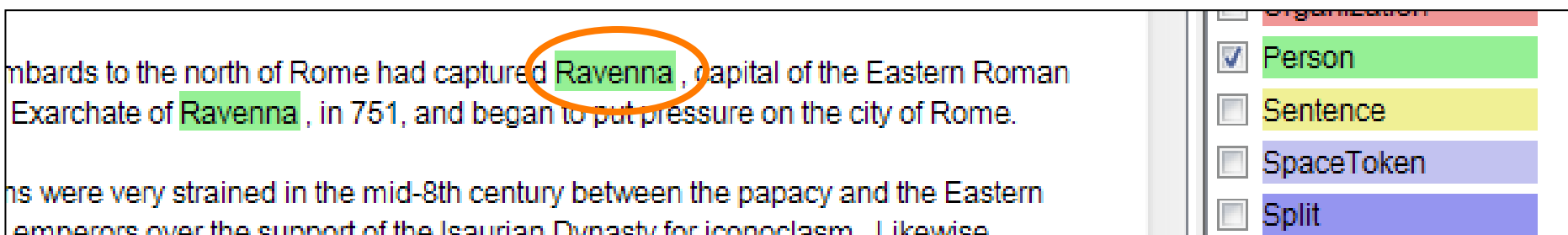
## Saving documents outside GATE

- 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 documents outside GATE

mbards to the north of Rome had captured **Ravenna**, capital of the Eastern Roman Exarchate of **Ravenna**, in 751, and began to put pressure on the city of Rome.

ns were very strained in the mid-8th century between the papacy and the Eastern emperors over the support of the Isaurian Dynasty for iconoclasm. Likewise

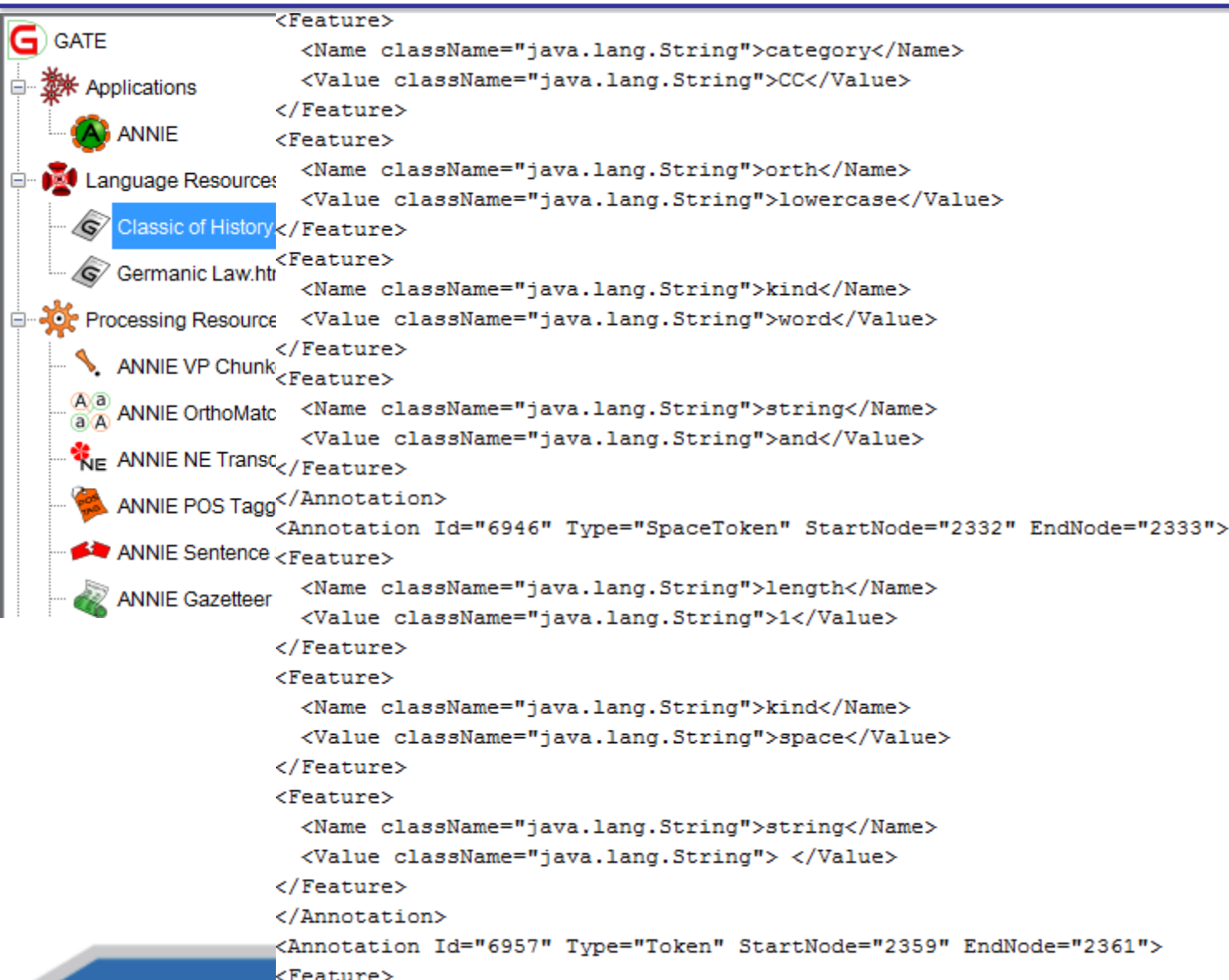


<input type="checkbox"/>	Organization
<input checked="" type="checkbox"/>	Person
<input type="checkbox"/>	Sentence
<input type="checkbox"/>	SpaceToken
<input type="checkbox"/>	Split

<Person> Ravenna </person>



# Saving as XML



The screenshot shows the GATE software interface. On the left, the 'Language Resources' pane lists several resources, with 'Classic of History' selected. The main area on the right displays the XML output of the document, showing the structure of the document with features and annotations.

```

<Feature>
  <Name className="java.lang.String">category</Name>
  <Value className="java.lang.String">CC</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">orth</Name>
  <Value className="java.lang.String">lowercase</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">kind</Name>
  <Value className="java.lang.String">word</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name>
  <Value className="java.lang.String">and</Value>
</Feature>
</Annotation>
<Annotation Id="6946" Type="SpaceToken" StartNode="2332" EndNode="2333">
  <Feature>
    <Name className="java.lang.String">length</Name>
    <Value className="java.lang.String">1</Value>
  </Feature>
  <Feature>
    <Name className="java.lang.String">kind</Name>
    <Value className="java.lang.String">space</Value>
  </Feature>
  <Feature>
    <Name className="java.lang.String">string</Name>
    <Value className="java.lang.String"> </Value>
  </Feature>
</Annotation>
<Annotation Id="6957" Type="Token" StartNode="2359" EndNode="2361">
  <Feature>

```

Save preserving format

```

</Split><p class="style1" align="left"><i><Token><Lookup>This</Lookup></Token> <Token>story</Token>
</Split><p class="style1" align="left"><Token>An</Token> <Token><Lookup><FirstPerson><Person>Arian<
</p><br/>
<p class="style1" align="left"><strong><Token><Lookup>II</Lookup></Token><Token><Split>.</Split></I
</Split><p class="style1" align="left"><i><Token><Lookup>This</Lookup></Token> <Token>is</Token> <T
</Split><p class="style1" align="left"><i>
</i><br/><Token>Now</Token> <Token>the</Token> <Token><Lookup><Lookup>one</Lookup></Lookup></Token>
</Split><p class="style1" align="left"><strong><Token><Lookup>III</Lookup></Token><Token><Split>.</
</Split><p class="style1" align="left"><i><Token><Lookup>This</Lookup></Token> <Token>is</Token> <T
</i></p><br/>
<p class="style1" align="left"><Token>If</Token> <Token>anyone</Token> <Token>shall</Token> <Token>
</blockquote><hr align="left"/><div class="style1" align="left"><blockquote>
<p class="style1"><Token></Token><Token>1</Token><Token></Token></Token> <Token><Unknown>Arians</Unknown>
</blockquote></div><blockquote class="style1">
<p class="style1" align="left"><Token></Token><Token>2</Token><Token></Token></Token> <Token><Lookup>That
</Split><p class="style1" align="left"><Token></Token><Token>3</Token><Token></Token></Token> <Token><Loc
</blockquote><hr align="left"/><div class="style1" align="left"><blockquote>&#160;</blockquote></div>
<blockquote class="style1">
<p class="style1" align="left"><Token><Lookup>From</Lookup></Token><Token>:</Token> <Person><Token>
</blockquote><div align="left"><div align="left"></div></div>
<p>&#160;</p><Split>
</Split><p><Token><Unknown>Introduction</Unknown></Token> <Token>and</Token> <Token>e-text</Token>
</blockquote></div></td></tr>

```

## References

- <http://gate.ac.uk/sale/tao/split.html>
- Introduction to GATE Developer [PDF document].  
Retrieved Web site: <http://gate.ac.uk>

You can download the program from:

- <http://gate.ac.uk/download/>