Practical NLP and Information Extraction with GATE

Dr. Diana Maynard

University of Sheffield, UK

What is text mining?

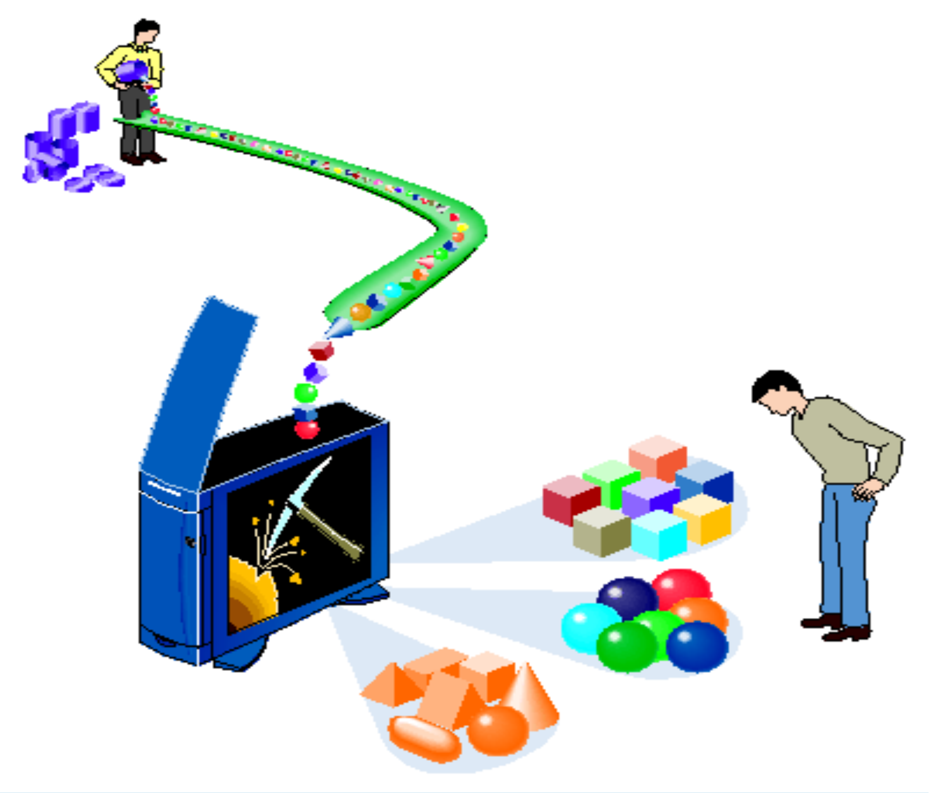
- Text Mining is the discovery of new, previously unknown information, by automatically extracting information from different textual resources.
- A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further
- Text mining lets you investigate what’s actually in a document or a collection of documents
- It lets us answer wh-questions: who, what, why, when, how, where, which?

Text Mining is not Data Mining

- Data mining is about using analytical techniques to find interesting patterns from large structured databases

Examples:

- using consumer purchasing patterns to predict which products to place close together on shelves in supermarkets
- analysing spending patterns on credit cards to detect fraudulent card use.



Text Mining is not Web Search

- Text mining is also different from traditional web search.
- In search, the user is typically looking for something that is already known and has been written by someone else.
- The problem lies in sifting through all the material that currently isn't relevant to your needs, in order to find the information that is.
- The solution often lies in better ways to ask the right question
- You can't ask Google to tell you:
  - How does the language used by Donald Trump differ from the language used by Hilary Clinton?
  - In which parts of the country did people talk more about the environment during the UK elections?
  - Which female MPs talked in the last 6 months about British hospitals with more than 100 deaths per month since 2010?

Basic text processing

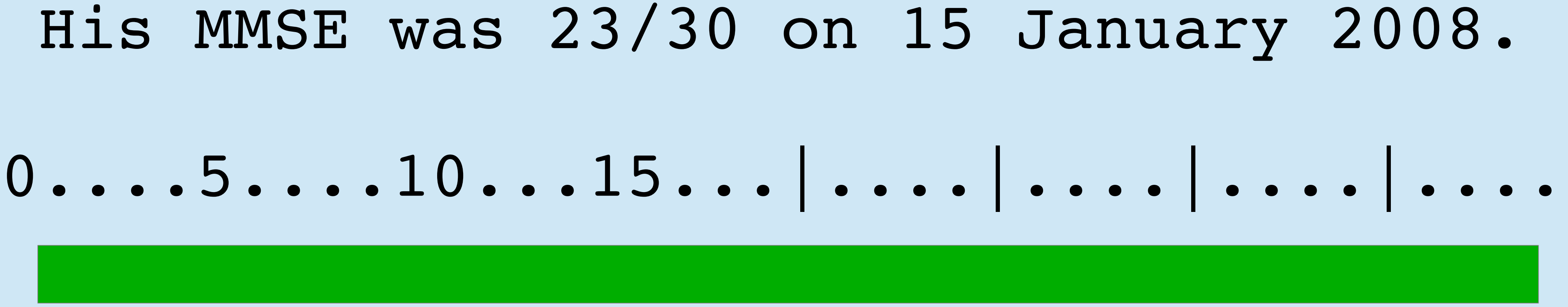
```
His MMSE was 23/30 on 15 January 2008.
```

(MMSE = Mini Mental State Examination)

Character offsets

```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

Sentences

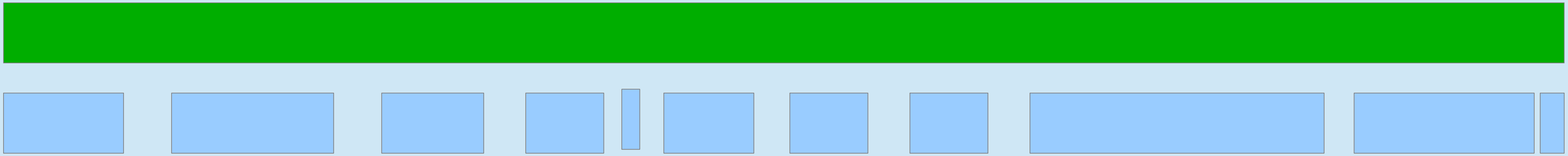


```
His MMSE was 23/30 on 15 January 2008.

0....5....10...15...|....|....|....|....
```

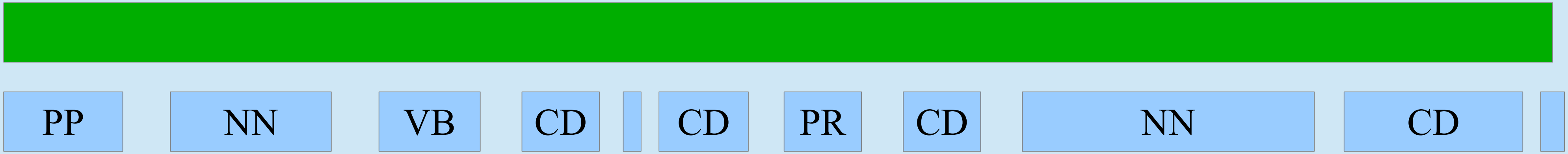
Tokens

```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



Part of speech categories

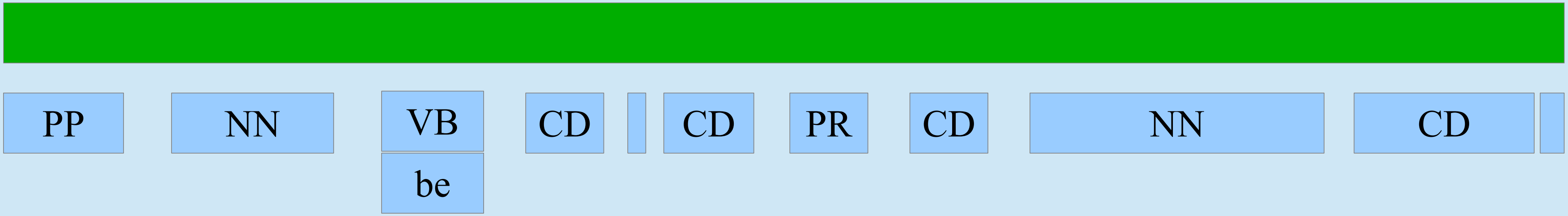
```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



PP NN VB CD CD PR CD NN CD

Morphological Analysis

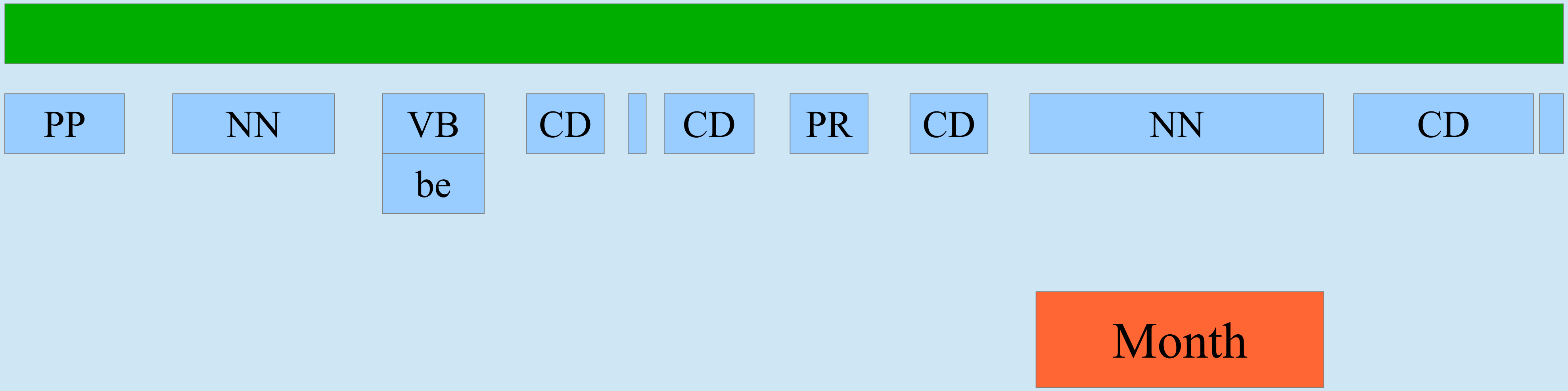
```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



PP NN VB be CD CD PR CD NN CD

Knowledge engineering: finding patterns

```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

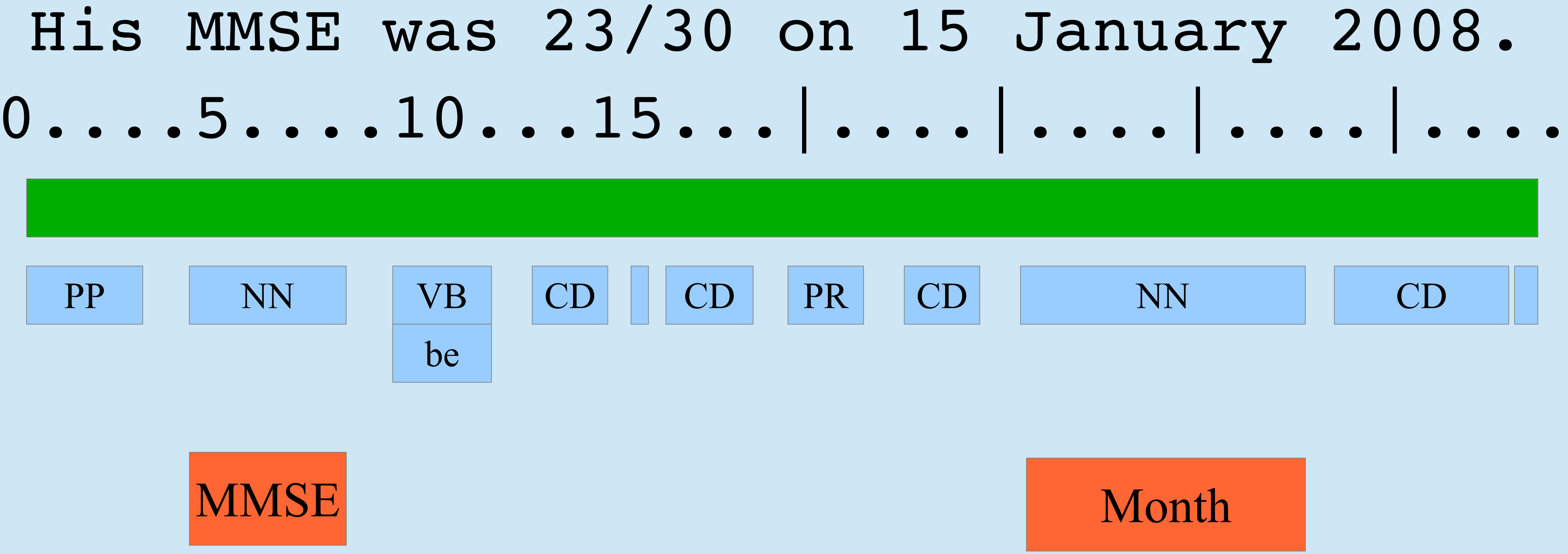


PP NN VB CD CD PR CD NN CD

be

Month

Knowledge engineering



```
 His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

PP

NN

VB

be

CD

CD

PR

CD

NN

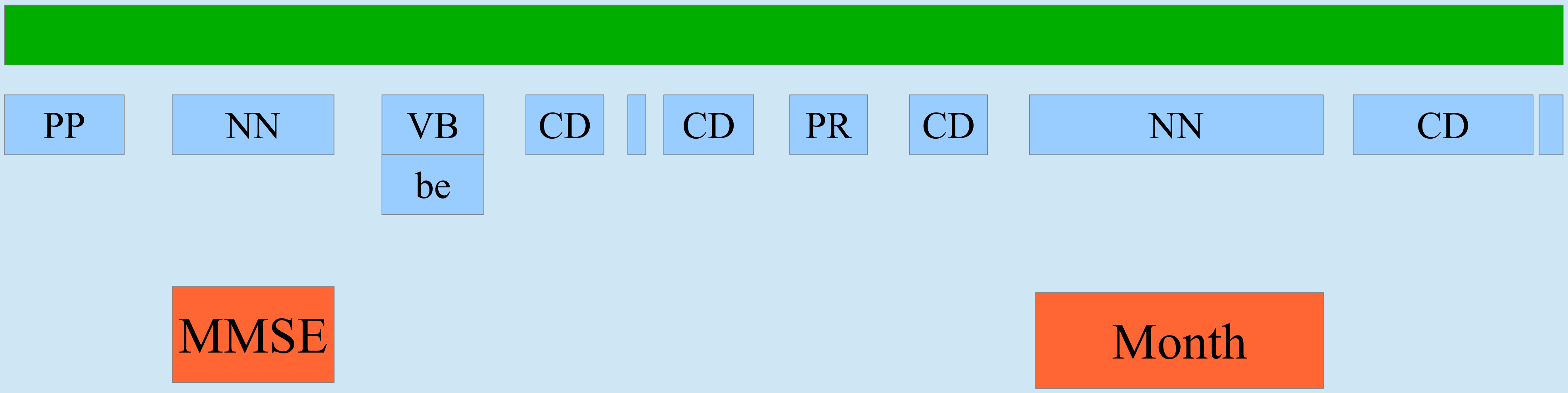
CD

MMSE

Month

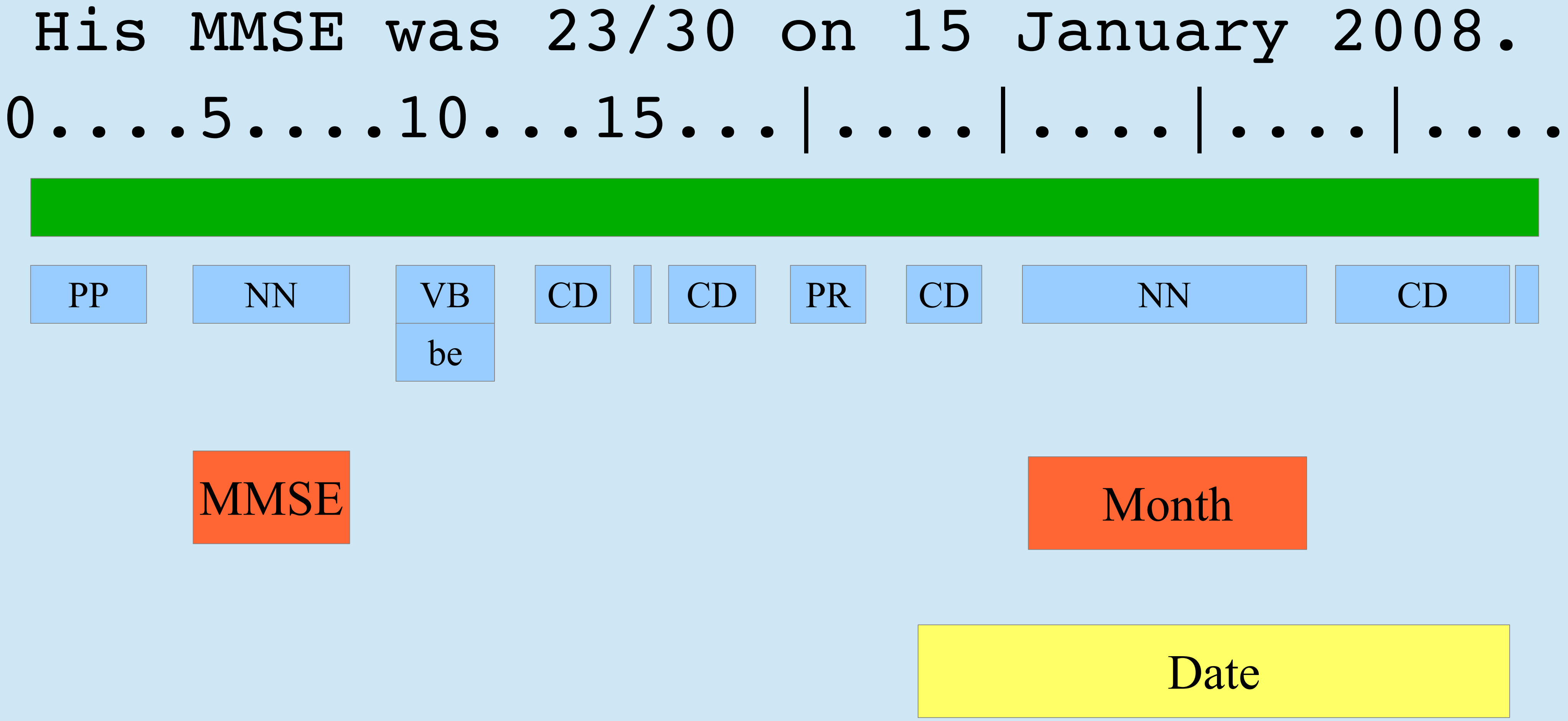
Knowledge engineering

```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```



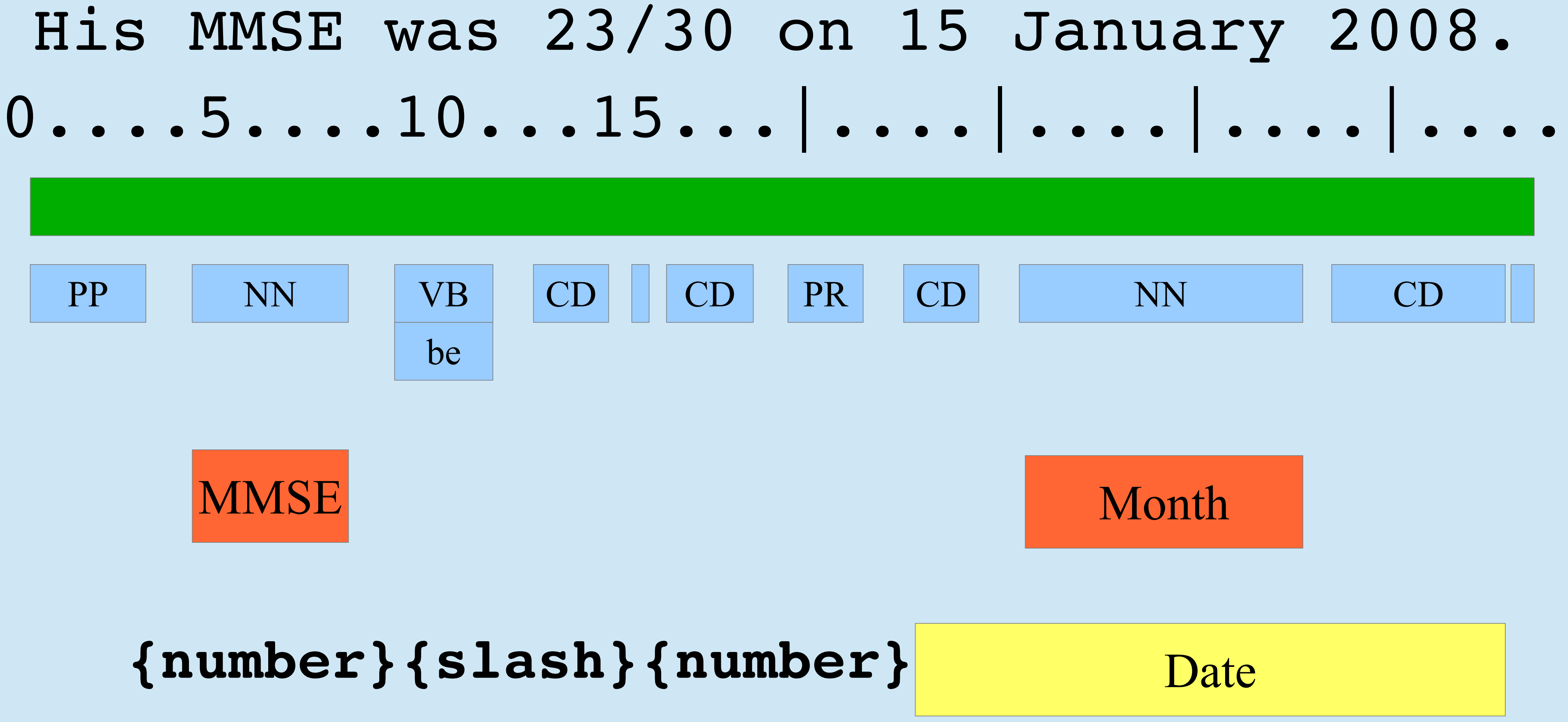
{number}{Month}{number}

Knowledge engineering



```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|....
```

Knowledge engineering



```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|.....
```

PP NN VB CD CD PR CD NN CD

be

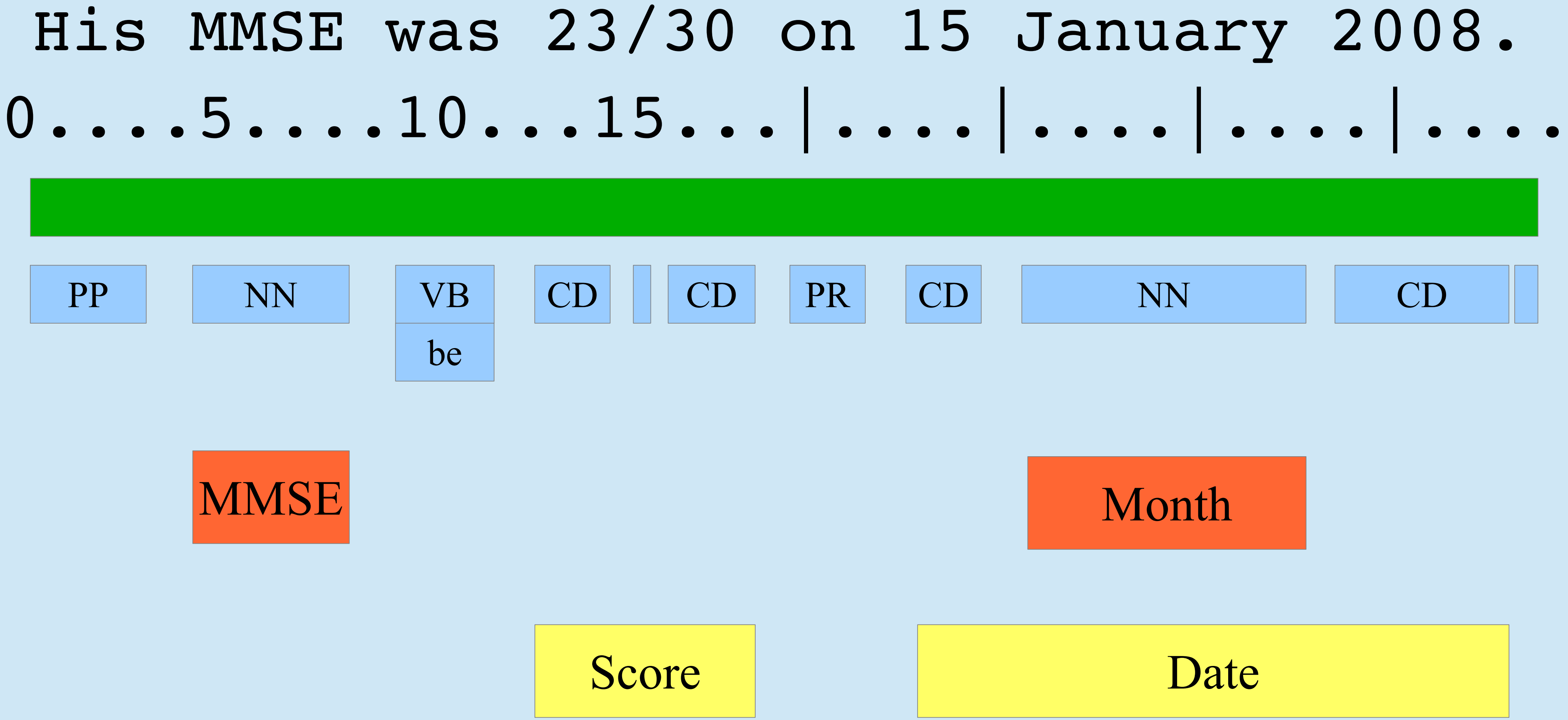
MMSE

Month

{number}{slash}{number}

Date

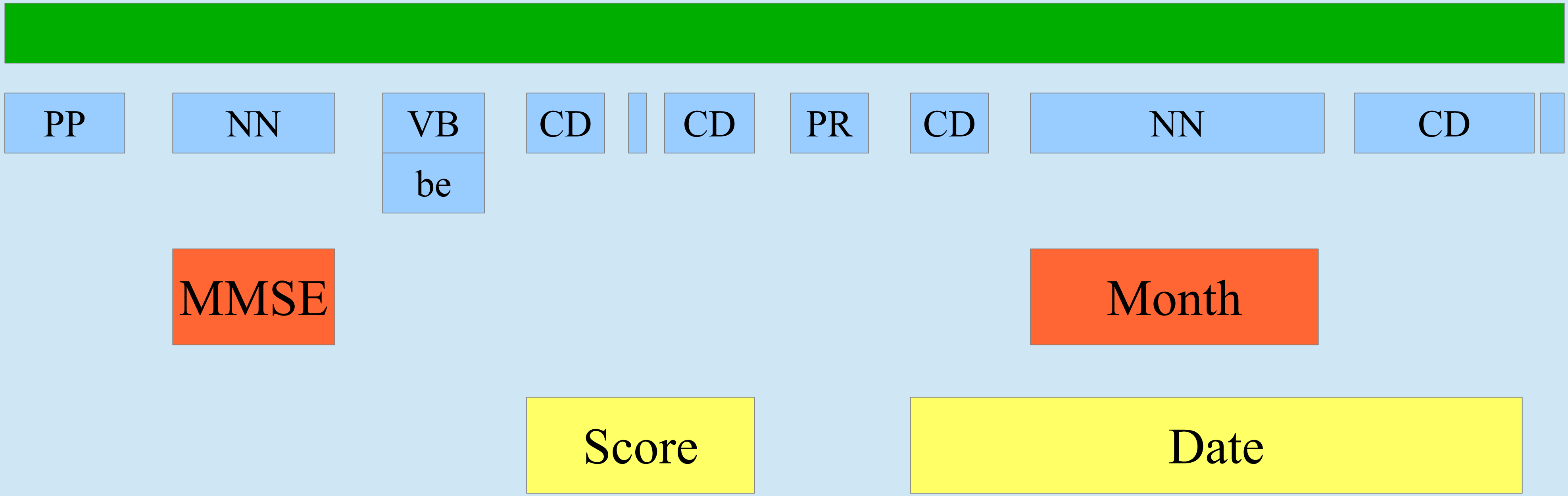
Knowledge engineering



```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|.....
```

Knowledge engineering

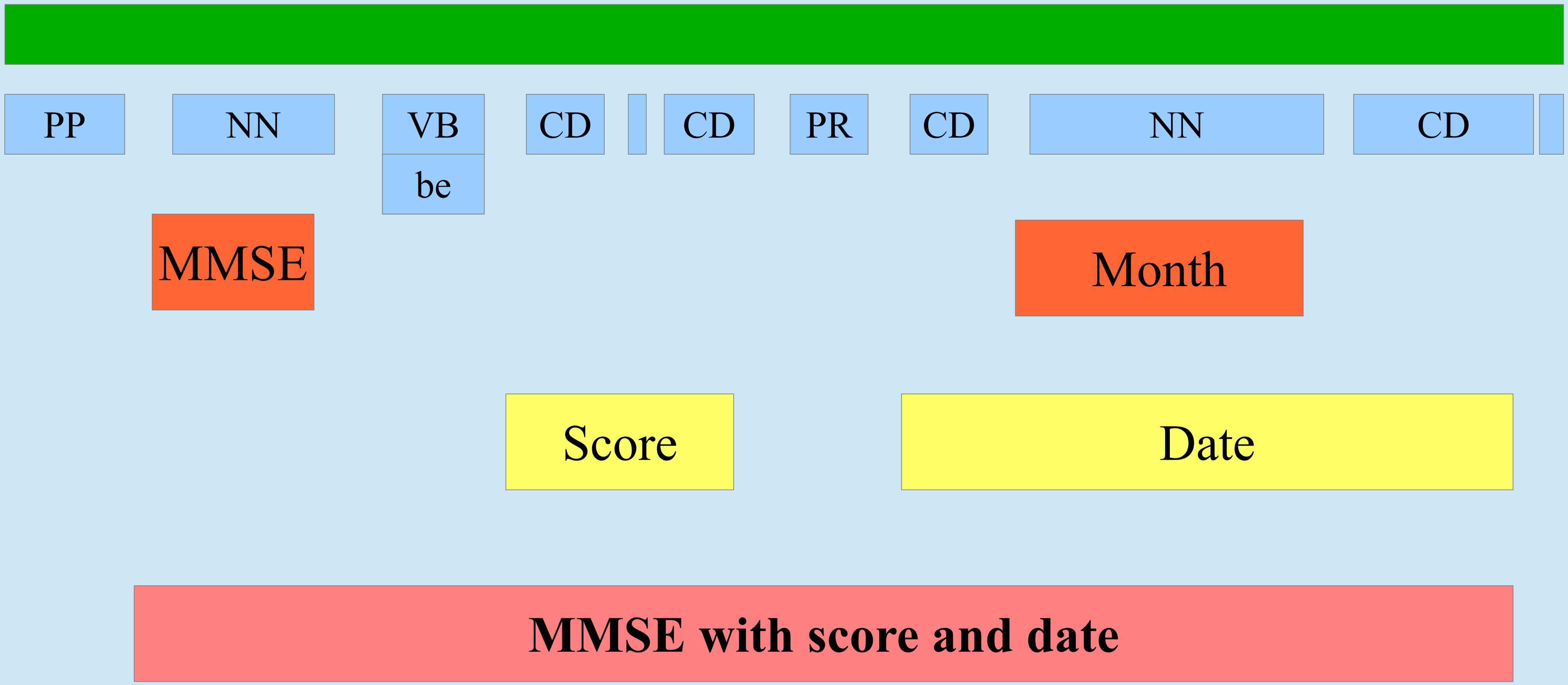
```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|.....
```



{MMSE}{BE}{Score}{?}{Date}

Knowledge engineering

```
His MMSE was 23/30 on 15 January 2008.
0....5....10...15...|....|....|....|.....
```



PP NN VB CD CD PR CD NN CD

be

MMSE

Month

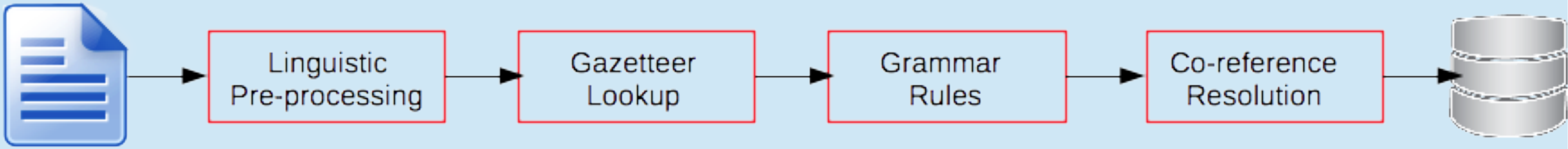
Score

Date

MMSE with score and date

Typical Information Extraction pipeline

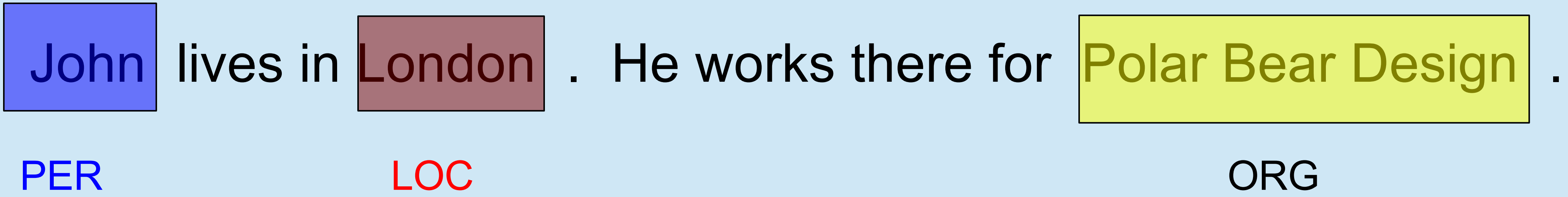
- Pre-processing (tokenisation, sentence splitting, morphological analysis, POS tagging)
- Entity finding (gazetteer lookup, NE grammars)
- Co-reference (alias finding, orthographic co-reference etc.)
- Export the results somewhere (database / XML / ontology)



Example of Information Extraction

John lives in London . He works there for Polar Bear Design .

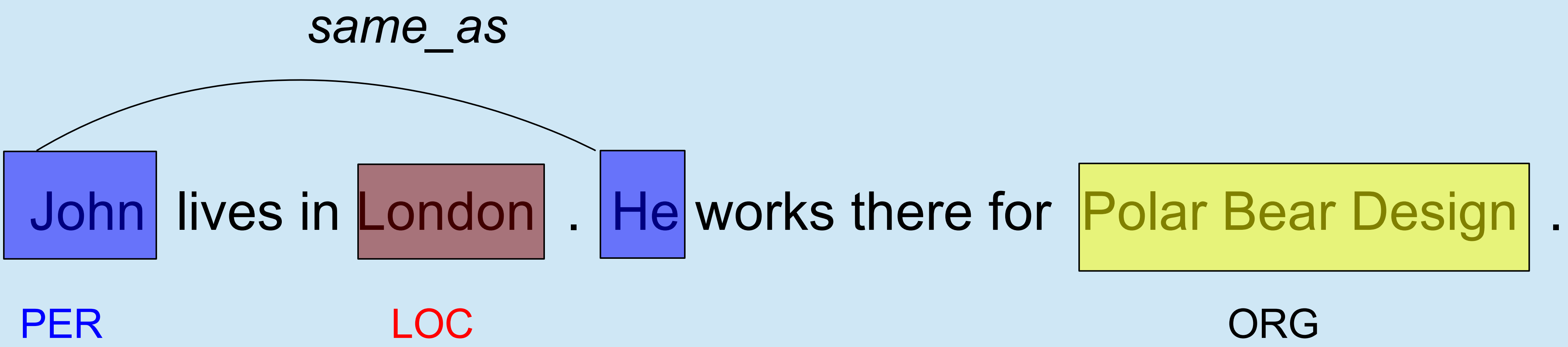
Basic Named Entity Recognition



John lives in London . He works there for Polar Bear Design .

PER LOC ORG

Co-reference



same_as

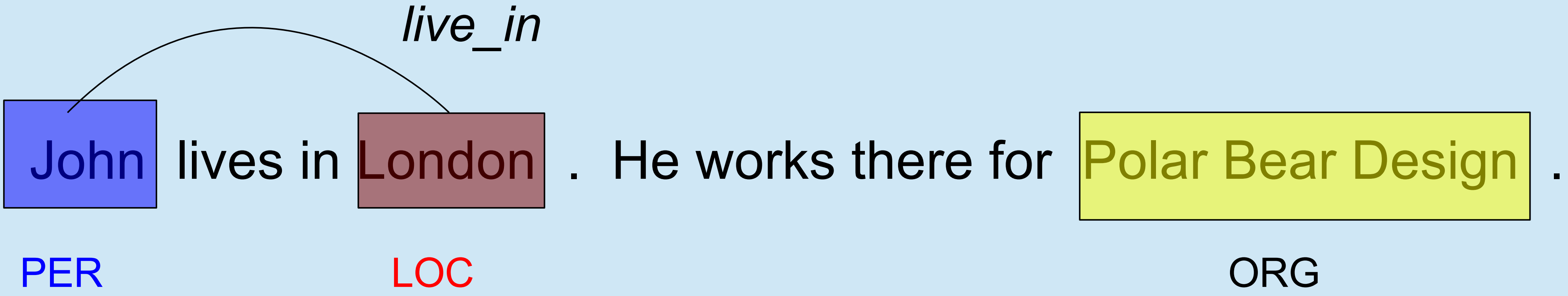
John lives in London . He works there for Polar Bear Design .

PER

LOC

ORG

Relations



live_in

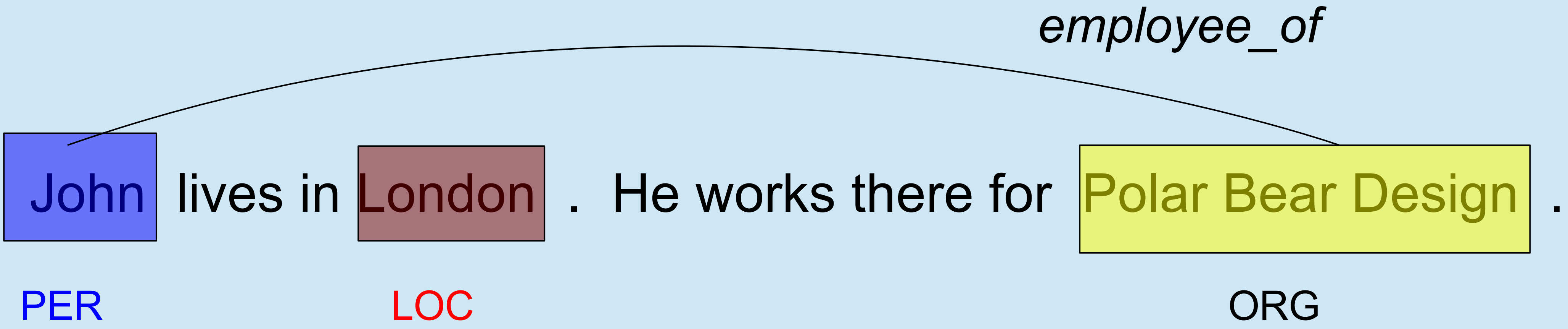
John lives in London . He works there for Polar Bear Design .

PER

LOC

ORG

Relations (2)

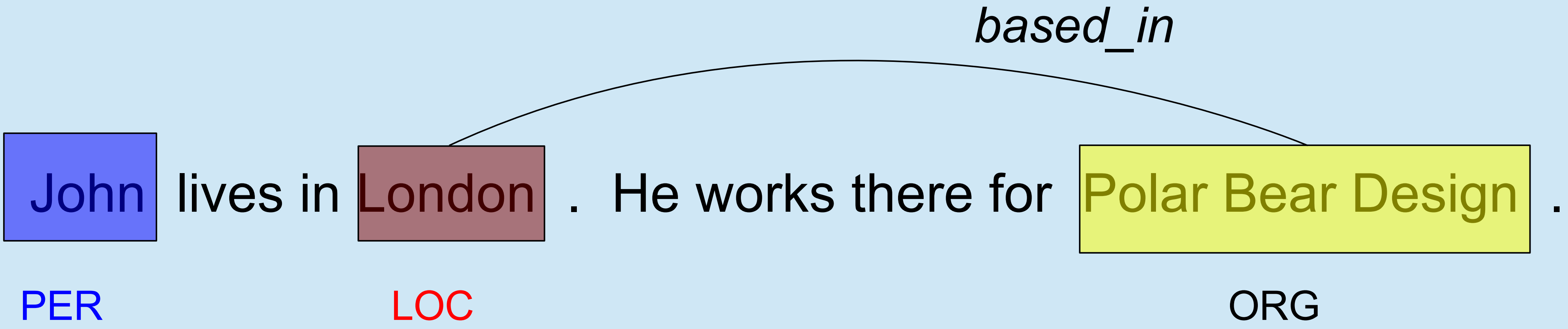


employee_of

John lives in London . He works there for Polar Bear Design .

PER LOC ORG

Relations (3)



based_in

John lives in London . He works there for Polar Bear Design .

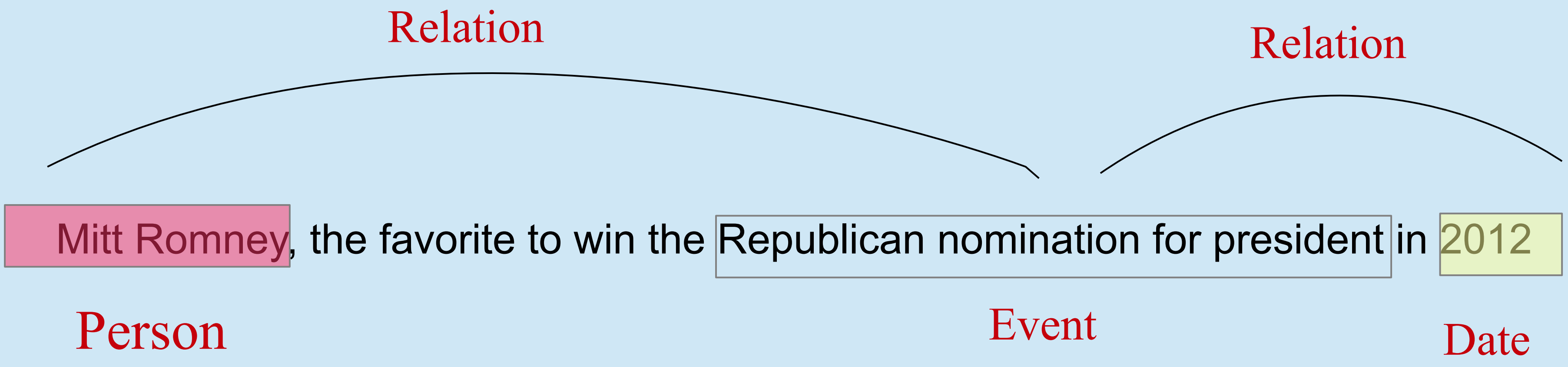
PER

LOC

ORG

What is Event Recognition?

- An event is an action or situation relevant to the domain expressed by some relation between entities or terms.
- It is always grounded in time, e.g. the performance of a band, an election, the death of a person



Relation

Relation

Mitt Romney, the favorite to win the Republican nomination for president in 2012

Person

Event

Date

Why are Entities and Events Useful?

- They can help answer the “Big 5” journalism questions (who, what, when, where, why)
- They can be used to categorise the texts in different ways
  - look at all texts about Donald Trump
  - They can be used as targets for opinion mining
  - find out what people think about Donald Trump
- When linked to an ontology and/or combined with other information, they can be used for reasoning about things not explicit in the text
  - seeing how opinions about different American presidents have changed over the years

GATE:

General Architecture for Text Engineering

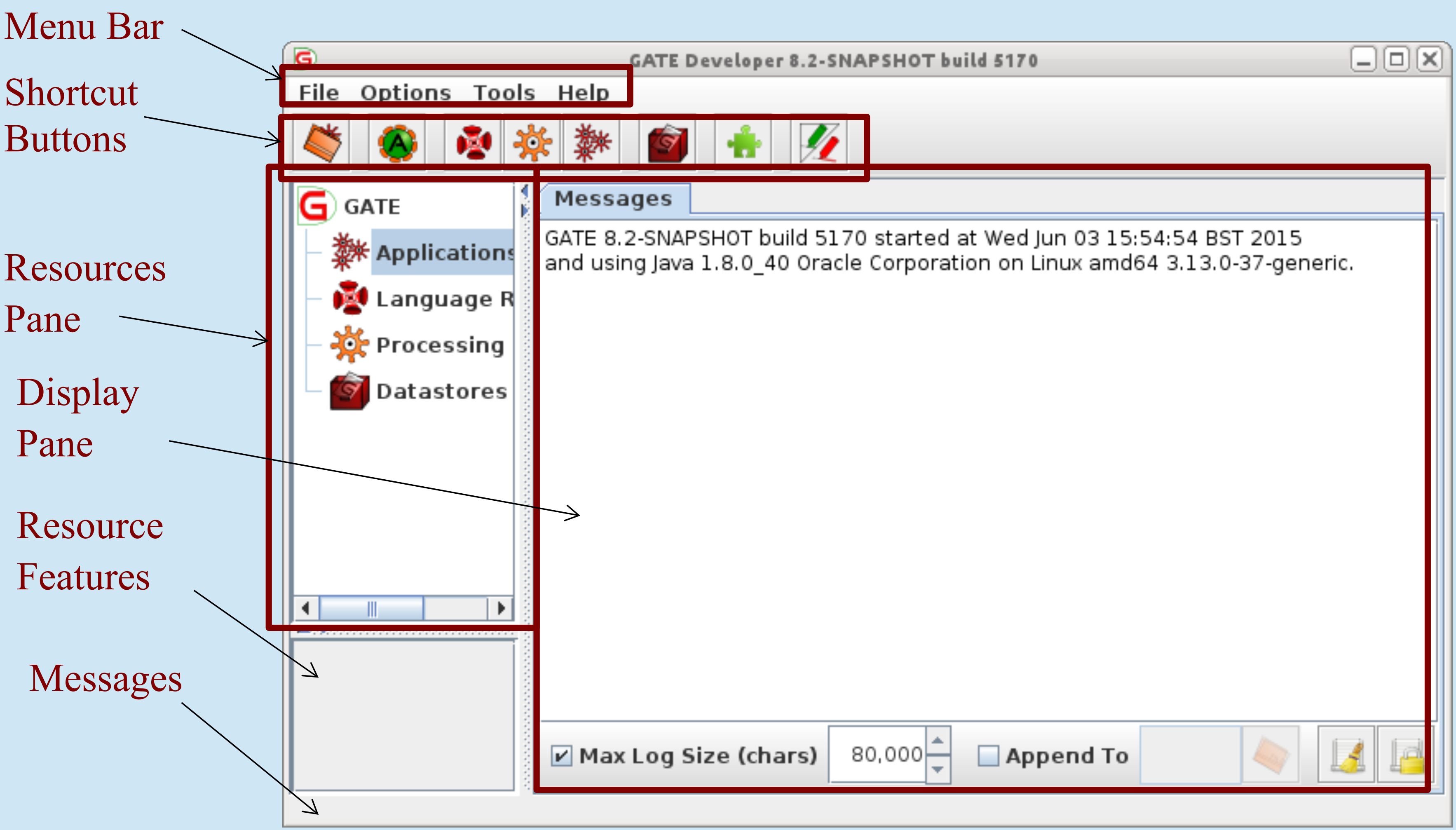
Why GATE?

- 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
- 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.

About this tutorial

- This tutorial will get you started with the GATE graphical user interface (GUI), also known as “GATE Developer”
- Everything you do in the GUI can be done via the API, but it’s easier to see what’s going on in the GUI
- It will be a hands-on session. You can try things out in GATE as the topics are presented.
- Things suggested for you to try yourself are in red.
- Download and install GATE 8.5.1 (if you haven’t already) from http://gate.ac.uk/download
- Start GATE on your computer (if you haven't already) by double clicking the icon

GATE GUI



GATE 8.2-SNAPSHOT build 5170 started at Wed Jun 03 15:54:54 BST 2015
and using Java 1.8.0_40 Oracle Corporation on Linux amd64 3.13.0-37-generic.

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

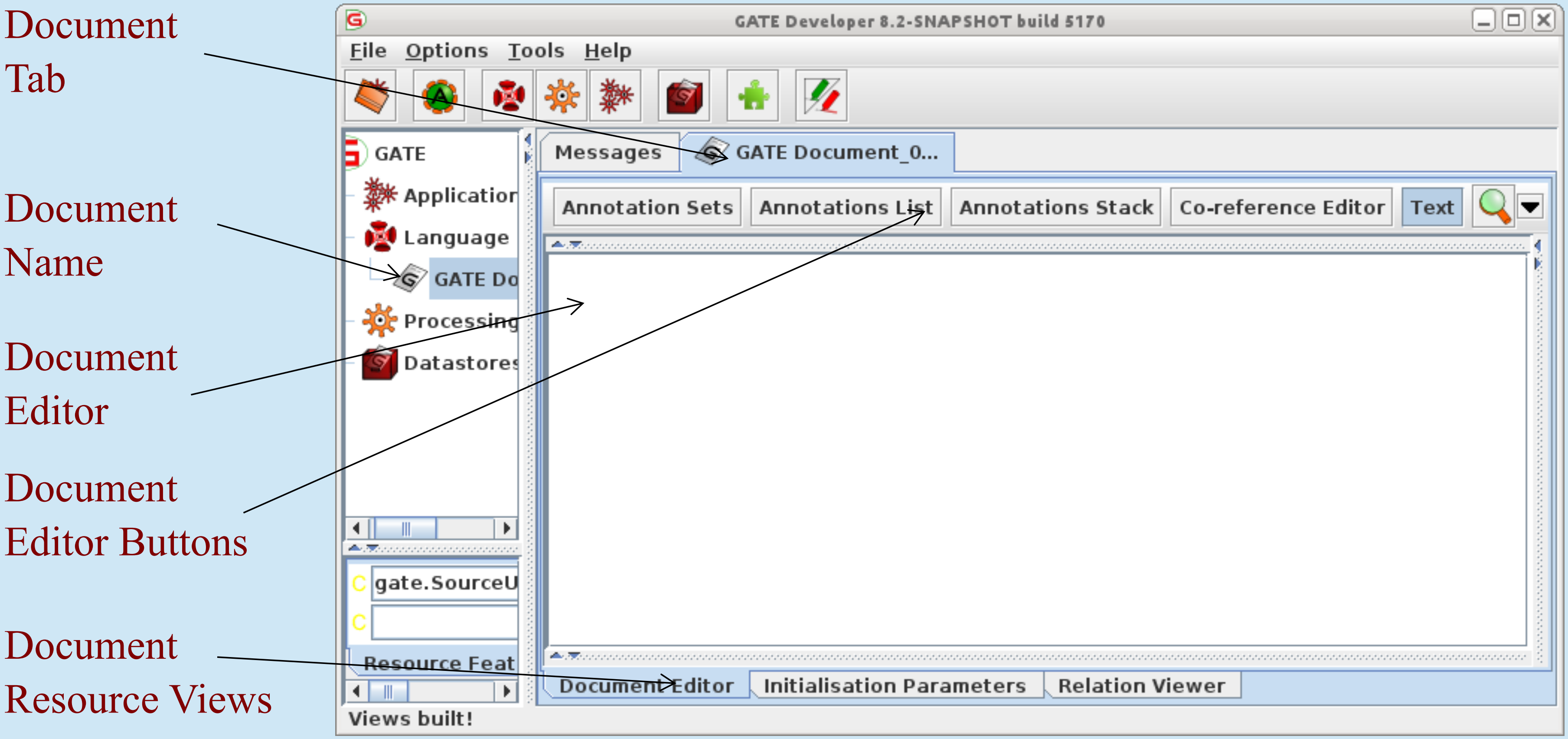
Displaying Resources

- 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 Resources 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.
- Now enter some text there.

Empty Document



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 an existing 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"

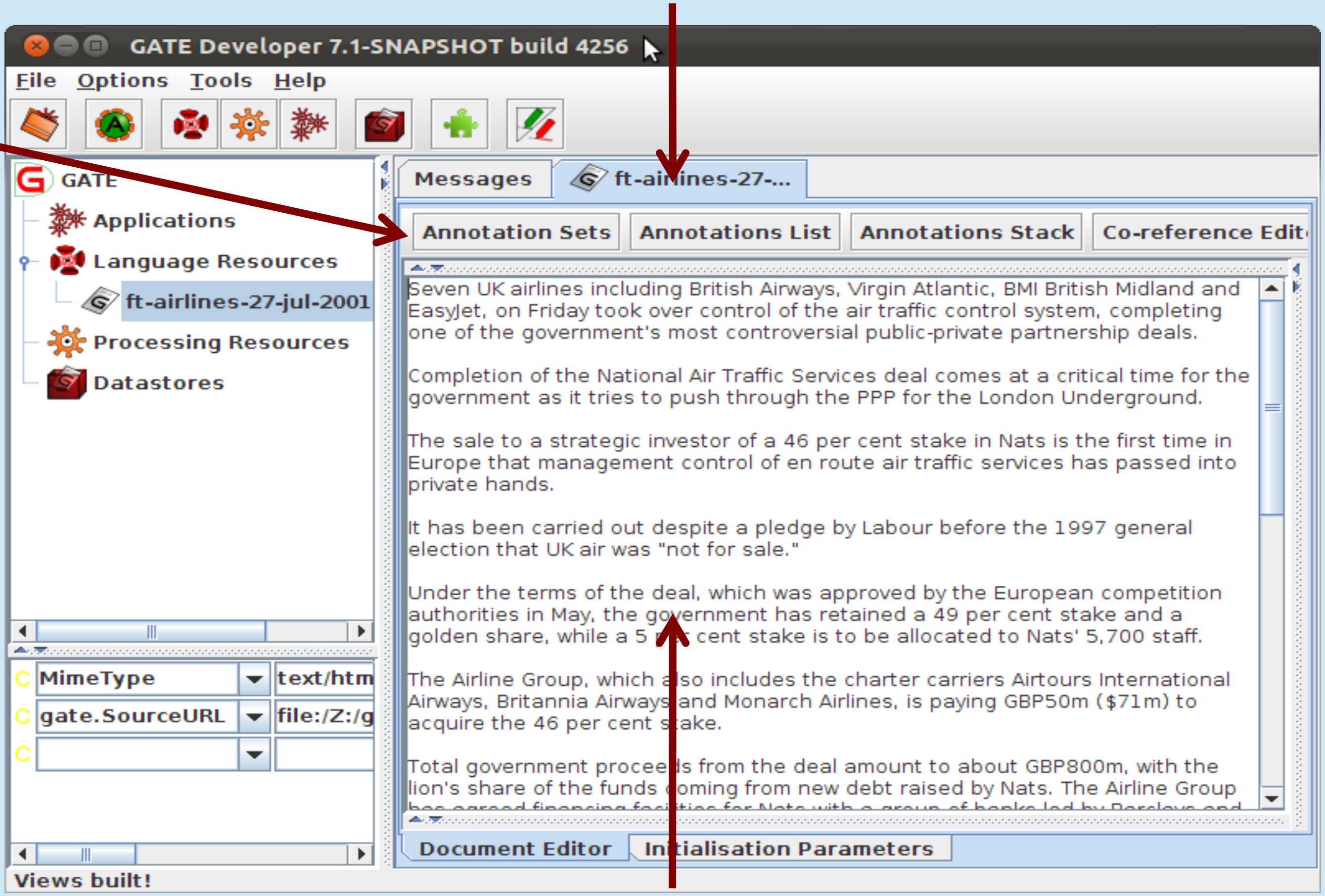
  OR

  - 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, e.g. http://news.bbc.co.uk**

Document viewer

Highlighted tab is the resource currently being viewed

Document viewer buttons



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

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

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

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

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

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

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

Document

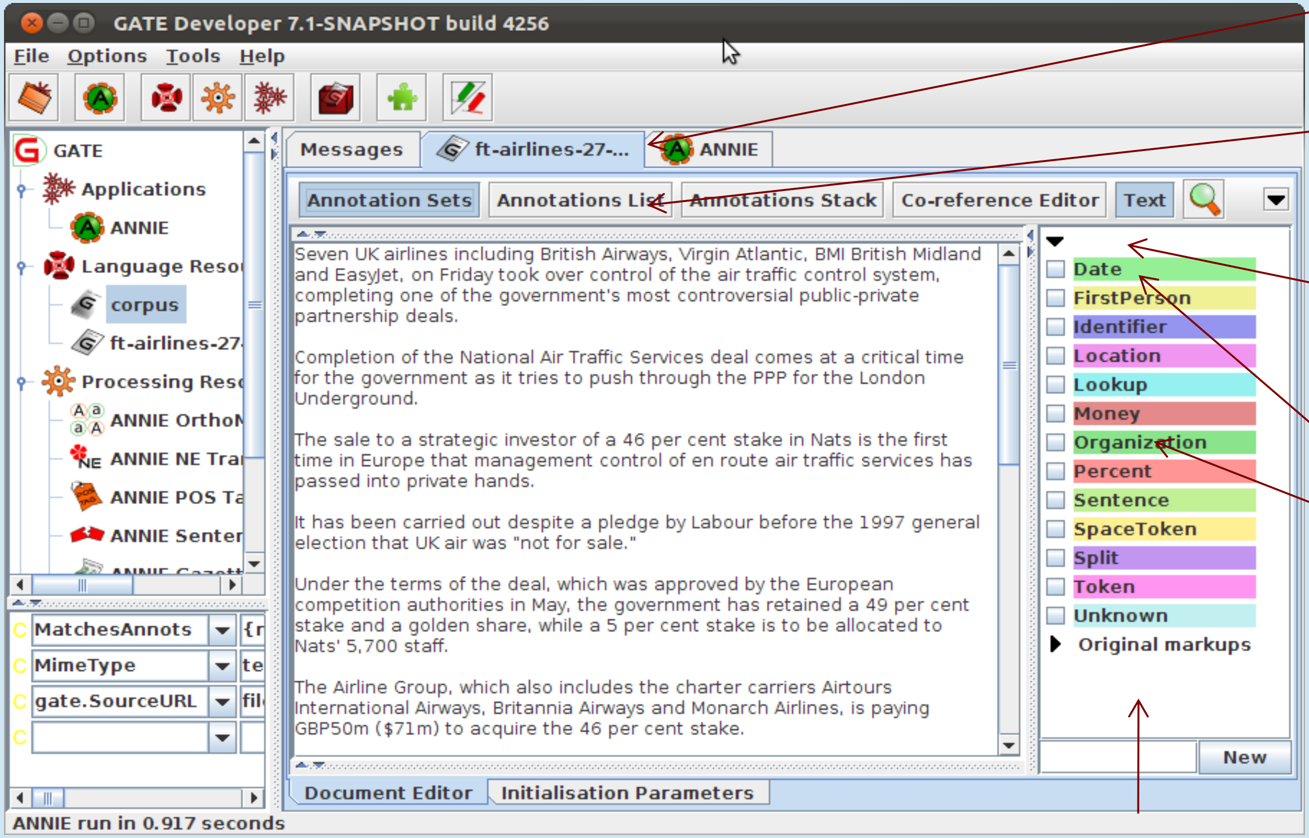
Annotations

- 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. gender=male

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 for the ft-airlines document you loaded

Annotation Sets



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

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

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

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

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

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

Tabs

Document Viewer Buttons

Default annotation set

Annotation types

Original markups annotation set

Viewing annotations

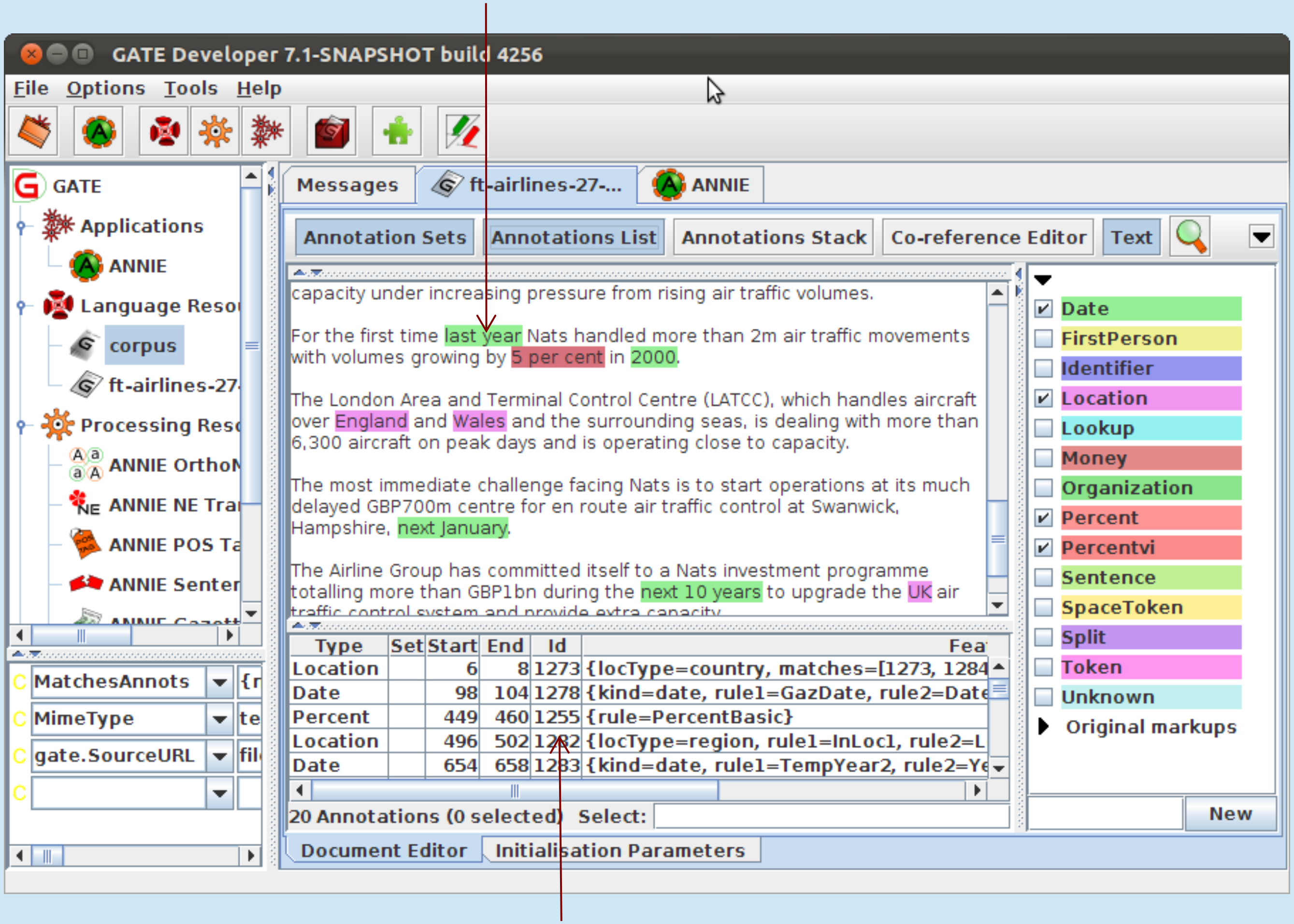
- Clicking on the Annotation Sets button opens a new pane on the right hand side inside the document view (Annotation Sets view)
- Default (unnamed) set contains some examples of annotations
- Click on the Annotation Set name (eg Key) to display the annotation types belonging to that set
- You should 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
- Try the Annotation Stack view

Annotations

Date annotation



capacity under increasing pressure from rising air traffic volumes.

For the first time last year Nats handled more than 2m air traffic movements with volumes growing by 5 per cent in 2000.

The London Area and Terminal Control Centre (LATCC), which handles aircraft over England and Wales and the surrounding seas, is dealing with more than 6,300 aircraft on peak days and is operating close to capacity.

The most immediate challenge facing Nats is to start operations at its much delayed GBP700m centre for en route air traffic control at Swanwick, Hampshire, next January.

The Airline Group has committed itself to a Nats investment programme totalling more than GBP1bn during the next 10 years to upgrade the UK air traffic control system and provide extra capacity.

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

20 Annotations (0 selected) Select:

- ☑ Date
- ☐ FirstPerson
- ☐ Identifier
- ☑ Location
- ☐ Lookup
- ☐ Money
- ☐ Organization
- ☑ Percent
- ☑ Percentvi
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Token
- ☐ Unknown
- ▶ Original markups

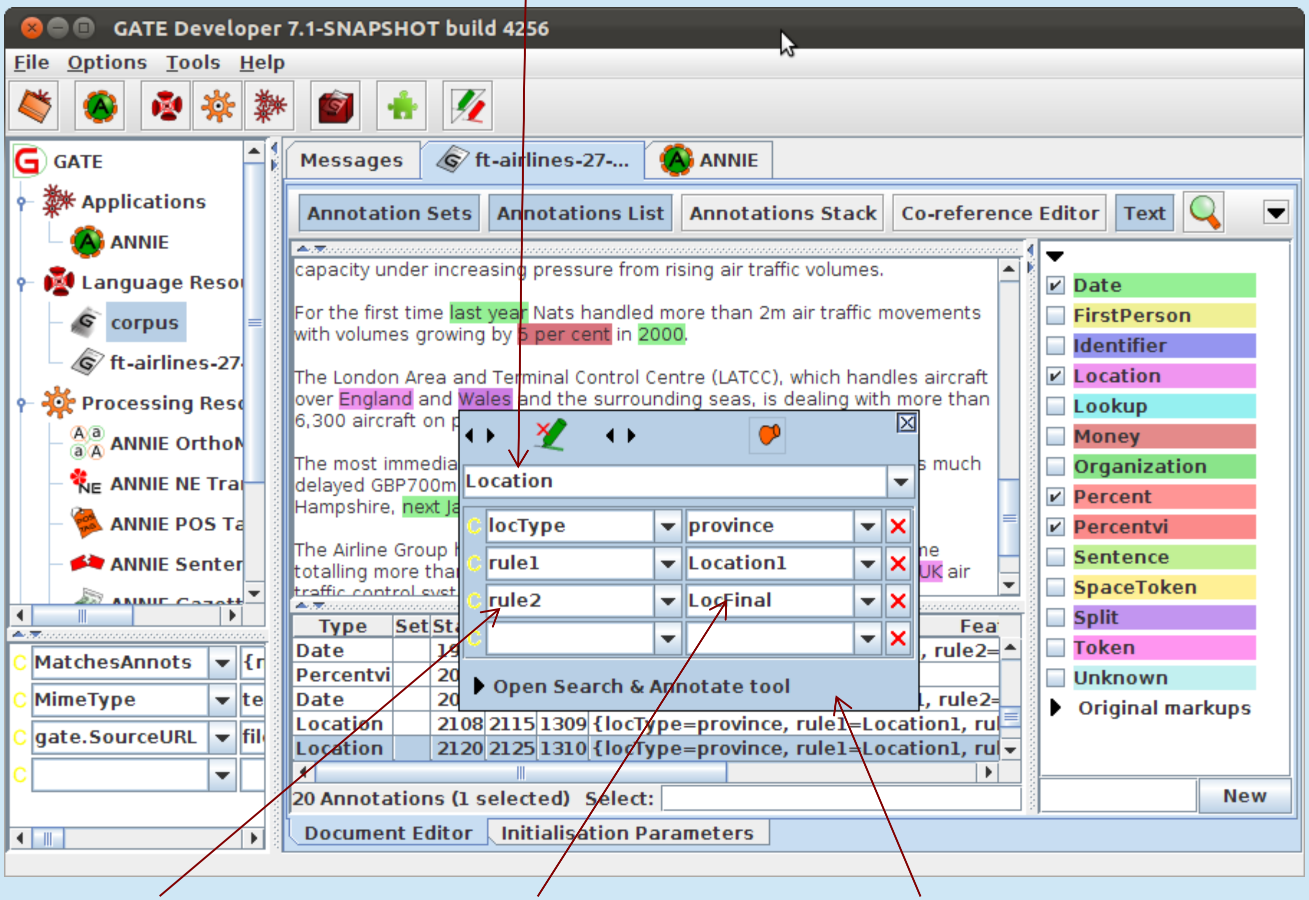
Annotations table

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



capacity under increasing pressure from rising air traffic volumes.

For the first time last year Nats handled more than 2m air traffic movements with volumes growing by 5 per cent in 2000.

The London Area and Terminal Control Centre (LATCC), which handles aircraft over England and Wales and the surrounding seas, is dealing with more than 6,300 aircraft on p

The most immedia… s much delayed GBP700m… Hampshire, next Ja…

The Airline Group … ne totalling more tha… UK air traffic control syst…

Annotation editor:

Location

locType	province
rule1	Location1
rule2	LocFinal

Type	Set	Sta			Fea
Date		19			, rule2=
Percentvi		20			
Date		20			, rule2=
Location		2108	2115	1309	{locType=province, rule1=Location1, ru
Location		2120	2125	1310	{locType=province, rule1=Location1, ru

20 Annotations (1 selected) Select:

MatchesAnnots	{r
MimeType	te
gate.SourceURL	fil

feature name

value

annotation editor

Creating a Corpus

- A corpus is a collection of documents.
- We tend to run applications on a corpus rather than on a document itself
- First close the documents you have loaded in GATE, just so we don’t get confused (right click and Close)
- Now create a new empty corpus
- Right click on Language Resources → New → GATE Corpus
- You can give the corpus a name, or use the default one

Populating a Corpus

- Sometimes there could be hundreds of documents in a corpus.
- Using the populate function means you can load lots of documents into the corpus in one go
- Right click on the name of your new corpus in the Resources pane and select “Populate”
- Select the name of the directory with your documents (hands-on/corpora/news-texts)
- All the documents will be loaded in one go
- View a document or the corpus by double clicking on it

Processing Resources

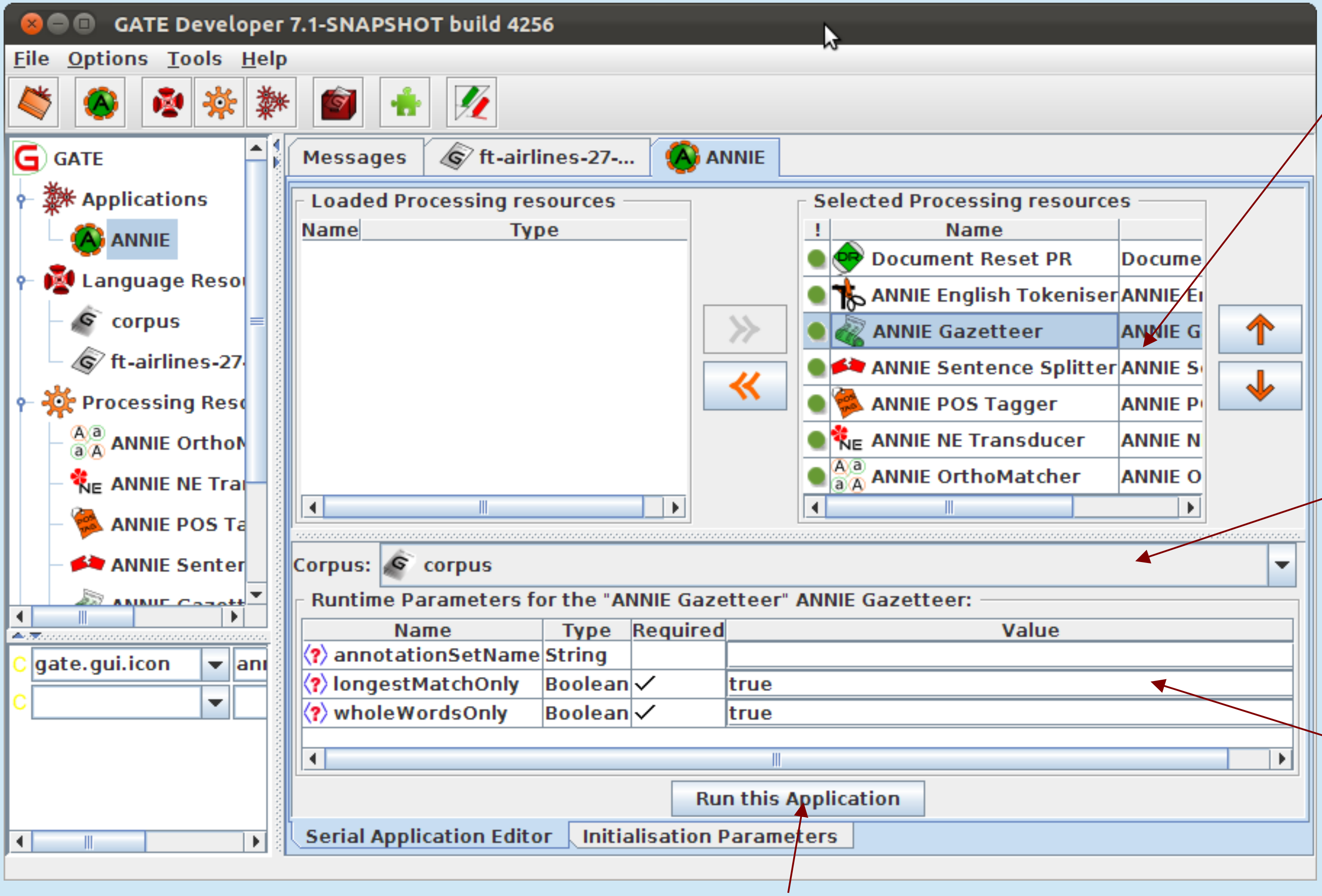
- Processing resources (PRs) are the tools that process and annotate text (text processing algorithms).
- An “application” or “pipeline” consists of any number of PRs, run sequentially over a corpus of documents
- ANNIE contains PRs for tokenisation, sentence splitting, POS tagging, Named Entity recognition etc.

Applications

Here's one we made earlier: ANNIE

- ANNIE is a ready-made collection of PRs that performs Information Extraction on unstructured text.
- Click the A icon from the top GATE menu OR Select File → Load ANNIE system
- View the ANNIE application by double clicking on it
- Run ANNIE on your corpus (select the corpus name and click “Run this application”)

Running an application



Loaded Processing resources

Name	Type

Selected Processing resources

!	Name	
●	Document Reset PR	Docume
●	ANNIE English Tokeniser	ANNIE E
●	ANNIE Gazetteer	ANNIE G
●	ANNIE Sentence Splitter	ANNIE S
●	ANNIE POS Tagger	ANNIE P
●	ANNIE NE Transducer	ANNIE N
●	ANNIE OrthoMatcher	ANNIE O

Corpus: corpus

Runtime Parameters for the "ANNIE Gazetteer" ANNIE Gazetteer:

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

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 or stack view too.
- Remember that not all the results will be perfect!

Plugins

- A plugin is a collection of PRs, and other resources bundled together.
    - Everything needed for IE in ANNIE is in the ANNIE plugin.
    - Everything needed for IE in French is in the lang_french 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

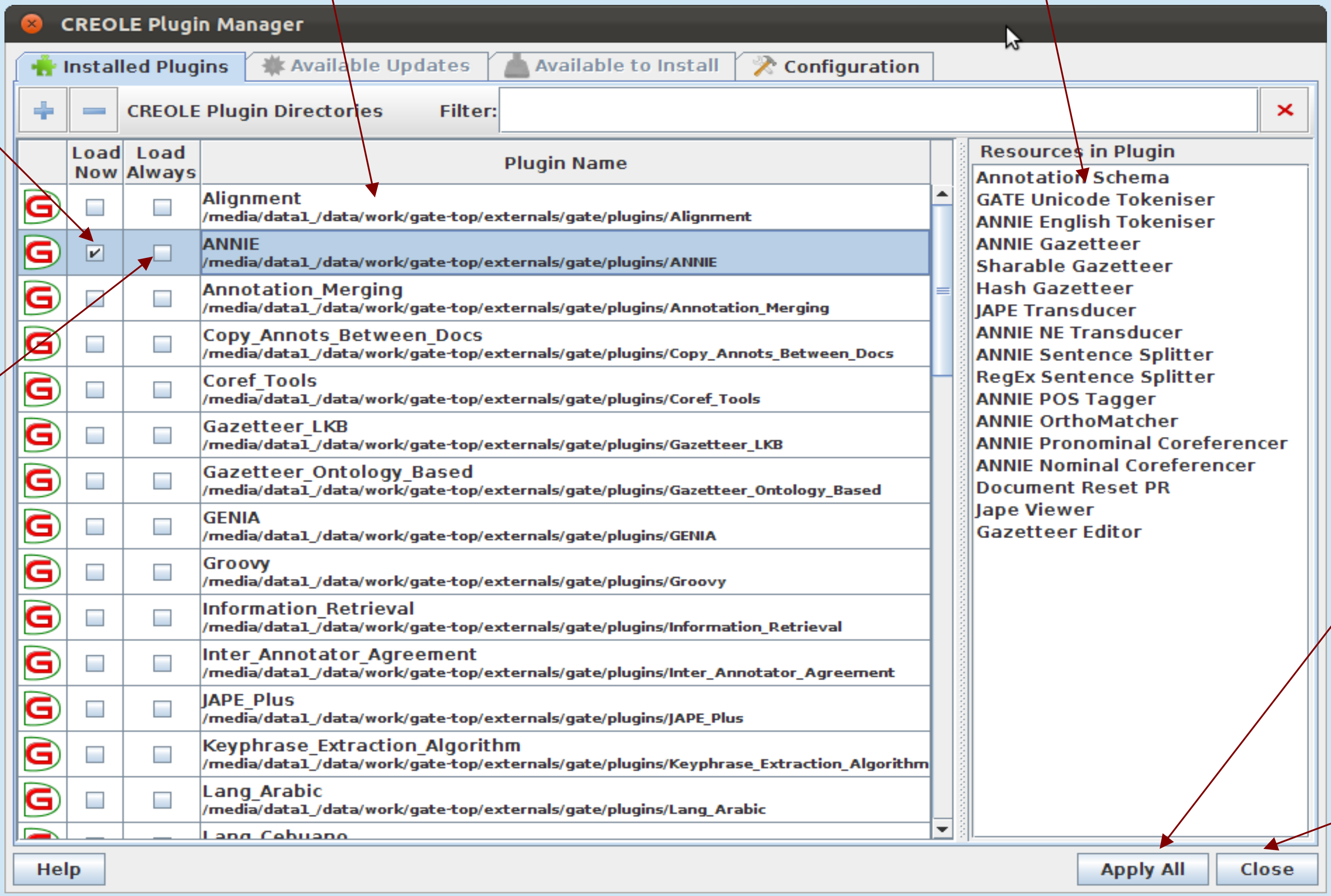
- Click the 🧩 icon on the top GATE menu to open the Plugin Manager [or go via File → Manage CREOLE Plugins]

Plugins

Load the plugin for this session only

List of available plugins

Resources in the selected plugin



	Load Now	Load Always	Plugin Name
G	□	□	Alignment /media/data1_/data/work/gate-top/externals/gate/plugins/Alignment
G	☑	□	ANNIE /media/data1_/data/work/gate-top/externals/gate/plugins/ANNIE
G	□	□	Annotation_Merging /media/data1_/data/work/gate-top/externals/gate/plugins/Annotation_Merging
G	□	□	Copy_Annots_Between_Docs /media/data1_/data/work/gate-top/externals/gate/plugins/Copy_Annots_Between_Docs
G	□	□	Coref_Tools /media/data1_/data/work/gate-top/externals/gate/plugins/Coref_Tools
G	□	□	Gazetteer_LKB /media/data1_/data/work/gate-top/externals/gate/plugins/Gazetteer_LKB
G	□	□	Gazetteer_Ontology_Based /media/data1_/data/work/gate-top/externals/gate/plugins/Gazetteer_Ontology_Based
G	□	□	GENIA /media/data1_/data/work/gate-top/externals/gate/plugins/GENIA
G	□	□	Groovy /media/data1_/data/work/gate-top/externals/gate/plugins/Groovy
G	□	□	Information_Retrieval /media/data1_/data/work/gate-top/externals/gate/plugins/Information_Retrieval
G	□	□	Inter_Annotator_Agreement /media/data1_/data/work/gate-top/externals/gate/plugins/Inter_Annotator_Agreement
G	□	□	JAPE_Plus /media/data1_/data/work/gate-top/externals/gate/plugins/JAPE_Plus
G	□	□	Keyphrase_Extraction_Algorithm /media/data1_/data/work/gate-top/externals/gate/plugins/Keyphrase_Extraction_Algorithm
G	□	□	Lang_Arabic /media/data1_/data/work/gate-top/externals/gate/plugins/Lang_Arabic
			Lang_Cebuano

Resources in Plugin

Annotation Schema
GATE Unicode Tokeniser
ANNIE English Tokeniser
ANNIE Gazetteer
Sharable Gazetteer
Hash Gazetteer
JAPE Transducer
ANNIE NE Transducer
ANNIE Sentence Splitter
RegEx Sentence Splitter
ANNIE POS Tagger
ANNIE OrthoMatcher
ANNIE Pronominal Coreferencer
ANNIE Nominal Coreferencer
Document Reset PR
Jape Viewer
Gazetteer Editor

Load the plugin everytime GATE starts

Apply all the settings

Close the plugins manager

Plugins

- Click on a plugin to see (on the RHS) the names of the resources it contains. Have a look at a few.
- Now load the Tools plugin by checking the relevant “Load Now” box for it
- Click “Apply All” to load the plugin
- Click “Close”
- Right click on Processing Resources to see which new PRs are now available

Adding a new PR

- 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.
- If you were looking closely, you'll have noticed that the Tools plugin you just loaded contains the ANNIE VP Chunker.
- Right click on Processing Resources and select "New" → "ANNIE VP Chunker"
- Leave all the default parameters set and click "OK"

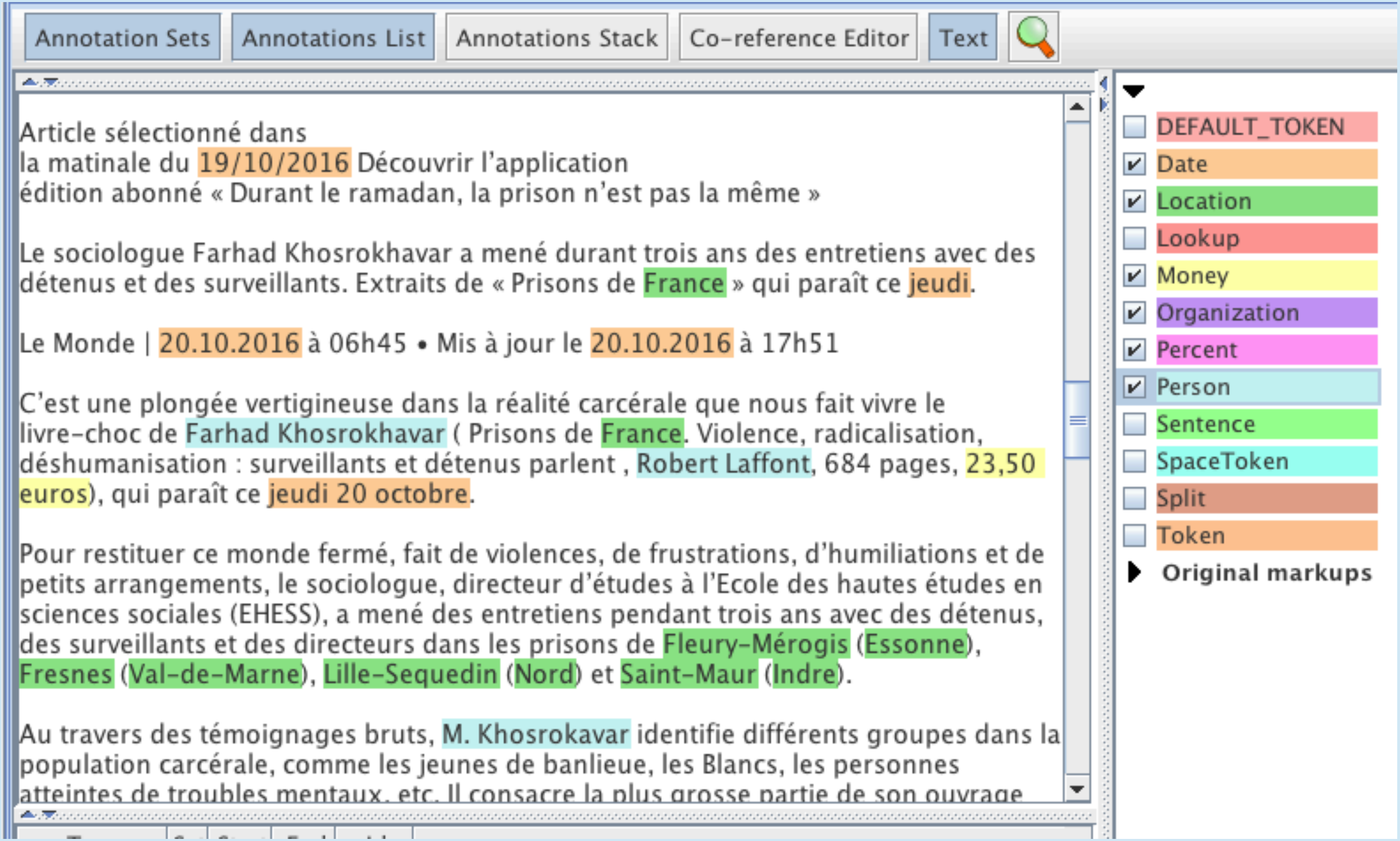
Adding a new PR (2)

- Now we need to add the new PR to the application.
- Double click on ANNIE.
- You'll see the ANNIE VP Chunker is in the list of loaded PRs. This means it's available in GATE, but isn't yet contained in the application.
- Add it to the application by selecting it and using the right arrow to transfer it.
- Now use the up arrow to move it to the right place in the application. It should go after (below) the POS tagger but before (above) the NE transducer.
- Run the application and view the results on the document.
- You should see a new annotation type "VG".

IE for other languages

- You can try out other applications in two ways
- Load a plugin that contains a ready-made application
- Click Applications -> Readymade applications
- Load some documents and run the application
- You can also just load a blank document and type some text in it
- If you do this, you need to right click on the document and select "New corpus with this document" first
- Run the new application on your corpus

NER in French



Article sélectionné dans
la matinale du 19/10/2016 Découvrir l'application
édition abonné « Durant le ramadan, la prison n'est pas la même »

Le sociologue Farhad Khosrokhavar a mené durant trois ans des entretiens avec des détenus et des surveillants. Extraits de « Prisons de France » qui paraît ce jeudi.

Le Monde | 20.10.2016 à 06h45 • Mis à jour le 20.10.2016 à 17h51

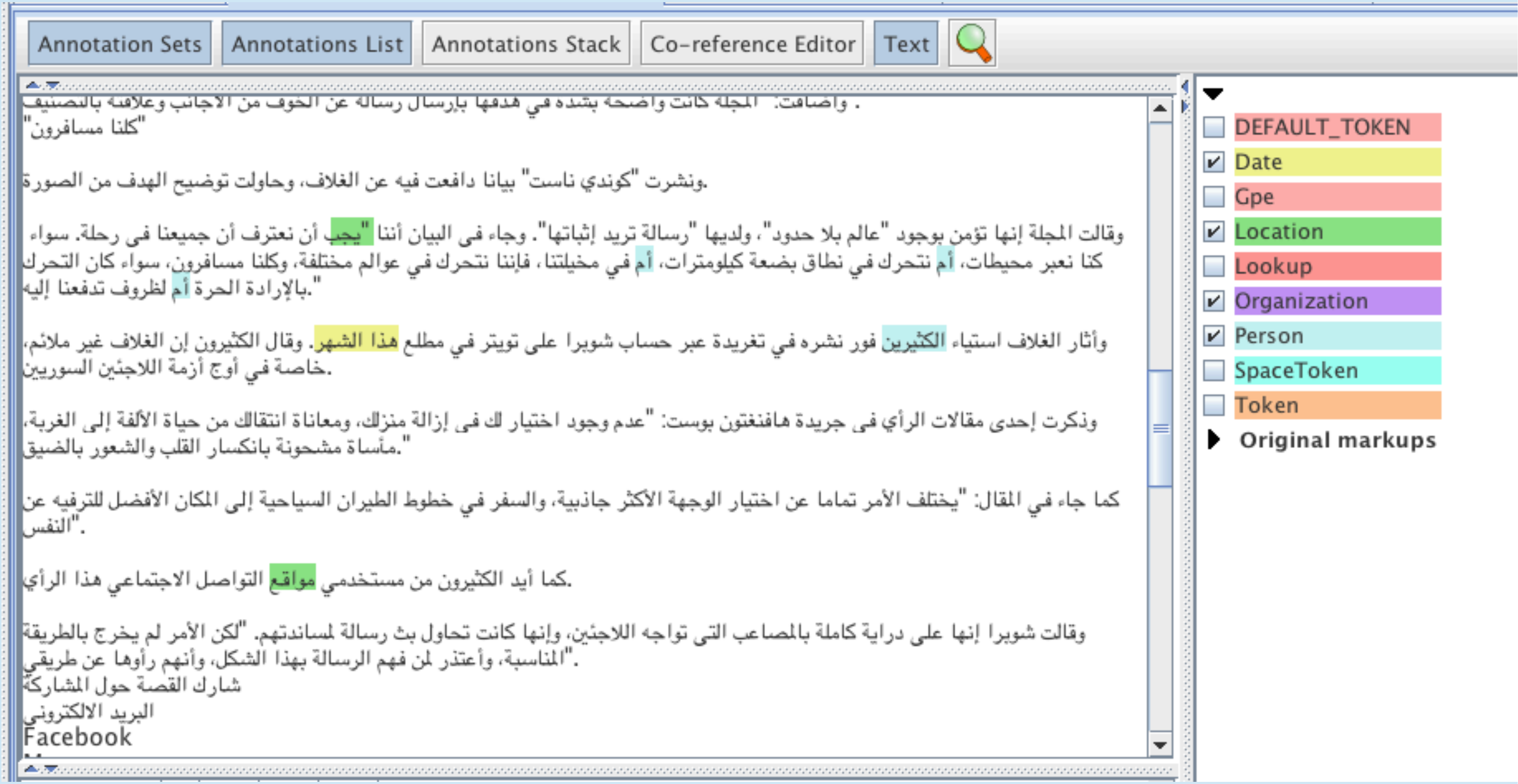
C'est une plongée vertigineuse dans la réalité carcérale que nous fait vivre le livre-choc de Farhad Khosrokhavar (Prisons de France. Violence, radicalisation, déshumanisation : surveillants et détenus parlent , Robert Laffont, 684 pages, 23,50 euros), qui paraît ce jeudi 20 octobre.

Pour restituer ce monde fermé, fait de violences, de frustrations, d'humiliations et de petits arrangements, le sociologue, directeur d'études à l'Ecole des hautes études en sciences sociales (EHESS), a mené des entretiens pendant trois ans avec des détenus, des surveillants et des directeurs dans les prisons de Fleury-Mérogis (Essonne), Fresnes (Val-de-Marne), Lille-Sequedin (Nord) et Saint-Maur (Indre).

Au travers des témoignages bruts, M. Khosrokavar identifie différents groupes dans la population carcérale, comme les jeunes de banlieue, les Blancs, les personnes atteintes de troubles mentaux, etc. Il consacre la plus grosse partie de son ouvrage

- ☐ DEFAULT_TOKEN
- ☑ Date
- ☑ Location
- ☐ Lookup
- ☑ Money
- ☑ Organization
- ☑ Percent
- ☑ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Token
- Original markups

NER in Arabic



وأضافت: "المجلة كانت واضحة بشدة في هدفها بإرسال رسالة عن الخوف من الأجانب وعلاقته بالتصنيف .
"كلنا مسافرون"

ونشرت "كوندي ناست" بيانا دافعت فيه عن الغلاف، وحاولت توضيح الهدف من الصورة.

وقالت المجلة إنها تؤمن بوجود "عالم بلا حدود"، ولديها "رسالة تريد إثباتها". وجاء فى البيان أننا "يجب أن نعترف أن جميعنا فى رحلة. سواء
كنا نعبر محيطات، أم نتحرك في نطاق بضعة كيلومترات، أم في مخيلتنا ، فإننا نتحرك في عوالم مختلفة، وكلنا مسافرون، سواء كان التحرك
بالإرادة الحرة أم لظروف تدفعنا إليه."

وأثار الغلاف استياء الكثيرين فور نشره في تغريدة عبر حساب شويرا على تويتر في مطلع هذا الشهر. وقال الكثيرون إن الغلاف غير ملائم،
خاصة في أوج أزمة اللاجئين السوريين.

وذكرت إحدى مقالات الرأي فى جريدة هافنغتون بوست: "عدم وجود اختيار لك فى إزالة منزلك، ومعاناة انتقالك من حياة الألفة إلى الغربة،
مأساة مشحونة بانكسار القلب والشعور بالضيق."

كما جاء في المقال: "يختلف الأمر تماما عن اختيار الوجهة الأكثر جاذبية، والسفر في خطوط الطيران السياحية إلى المكان الأفضل للترفيه عن
النفس".

كما أيد الكثيرون من مستخدمي مواقع التواصل الاجتماعي هذا الرأي.

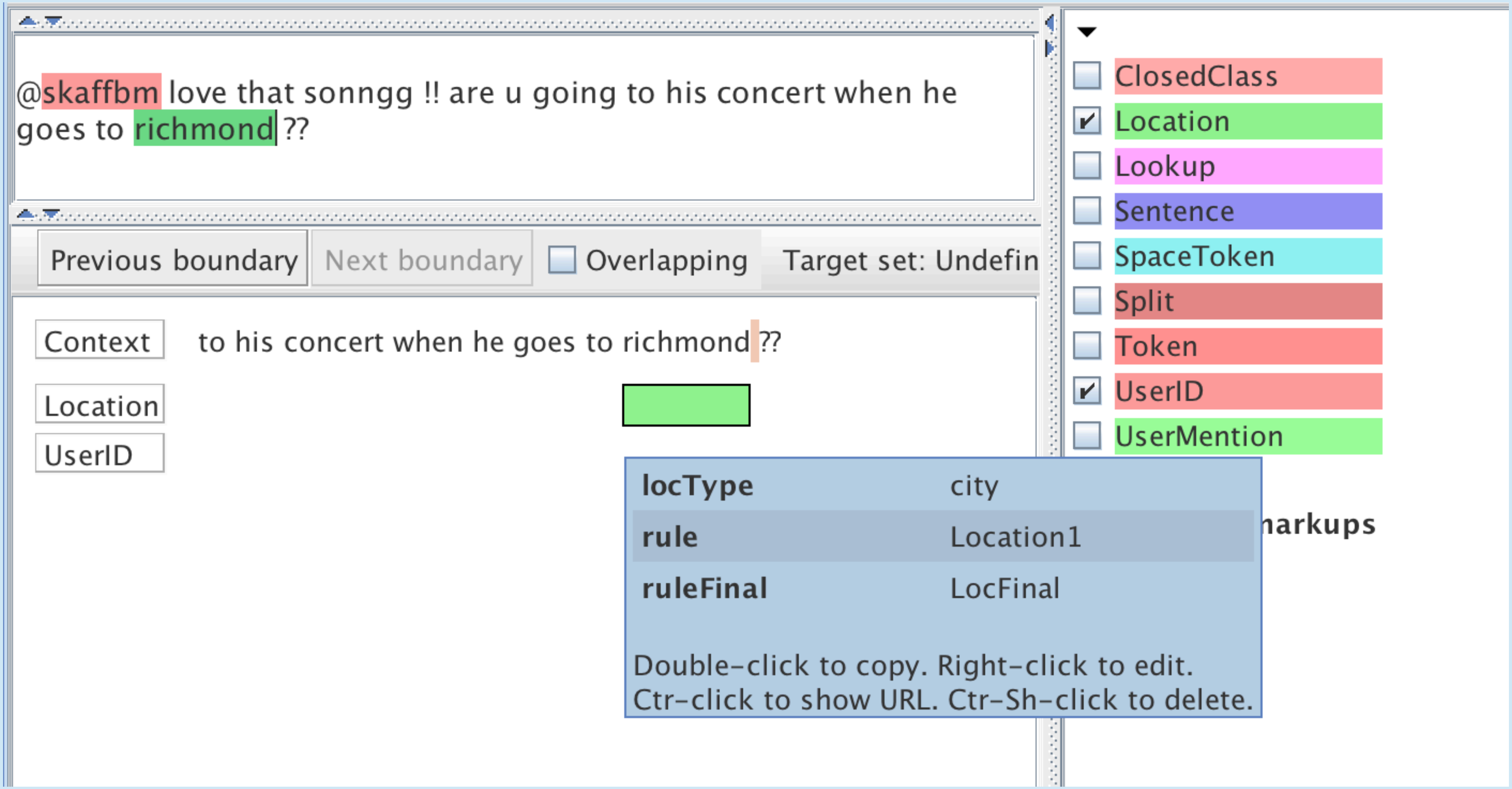
وقالت شويرا إنها على دراية كاملة بالمصاعب التى تواجه اللاجئين، وإنها كانت تحاول بث رسالة لمساندتهم. "لكن الأمر لم يخرج بالطريقة
المناسبة، وأعتذر لمن فهم الرسالة بهذا الشكل، وأنهم رأوها عن طريقي".
شارك القصة حول المشاركة
البريد الالكتروني
Facebook

☐ DEFAULT_TOKEN
☑ Date
☐ Gpe
☑ Location
☐ Lookup
☑ Organization
☑ Person
☐ SpaceToken
☐ Token
Original markups

Hands-on with TwitIE

- TwitIE is a version of ANNIE that's been retrained for tweets
- Load the TWITIE plugin (green jigsaw icon)
- Now right-click on "Applications", select "Ready-made applications" and "TwitIE"
- Create a new corpus, name it "Tweets"
- Right-click on the corpus and select "populate from Twitter JSON", selecting the file hands-on-materials/corpora/energy-tweets.json
- Once loaded, double click on TwitIE to open it, and then select "Run this application" (make sure the tweets corpus is chosen)
- Look at the different annotations in the default annotation set
- To see Tokens in hashtags, use the Annotation Stack view

Analysing tweets



@skaffbm love that sonngg !! are u going to his concert when he goes to richmond ??

to his concert when he goes to richmond ??

locType	city
rule	Location1
ruleFinal	LocFinal

Double-click to copy. Right-click to edit.
Ctr-click to show URL. Ctr-Sh-click to delete.

☐ ClosedClass
☑ Location
☐ Lookup
☐ Sentence
☐ SpaceToken
☐ Split
☐ Token
☑ UserID
☐ UserMention