



---

# Module 2: Introduction to IE and ANNIE



# About this tutorial

---

This tutorial comprises the following topics:

- Introduction to IE
- ANNIE
- Multilingual tools in GATE
- Evaluation and Corpus Quality Assurance

In Module 3, you'll learn how to use JAPE, the pattern matching language that many PRs use



# Tutorial outline

09:45 – 11:15

- What is information extraction?
- Examples of IE systems
- ANNIE
- Basic lexico-syntactic PRs

11:15 – 11:45

BREAK

11:45 – 13:15

- Gazetteers, transducers, coreference
- Modifying ANNIE
- Multilingual IE

13:15 – 14.15

LUNCH

14:15 – 15:45

- Evaluation
- Annotation Diff
- Corpus Quality Assurance

15:45 – 16:15

COFFEE

16:15 – 17:15

INVITED TALK – “Text mining Electronic Health Records for Research - Common Issues and Strategies ” - Richard Jackson

19:00 – 21:30

Social event: drinks reception and barbeque



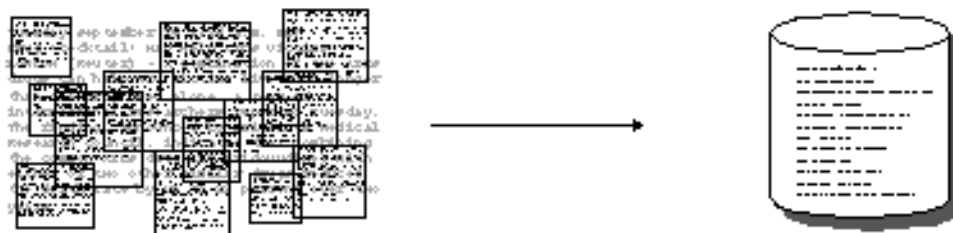
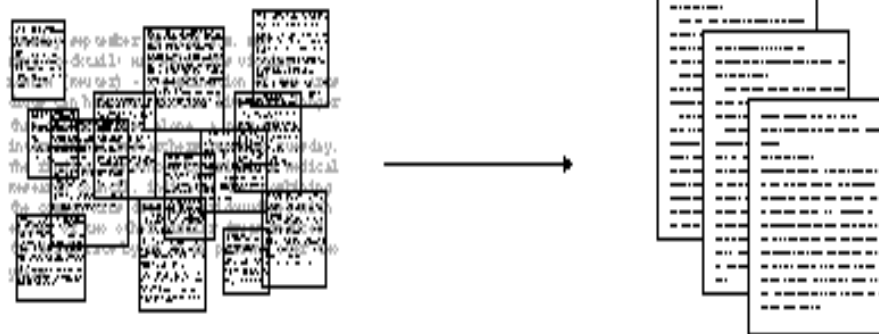
---

# What is information extraction?

# IE is not IR

GATE

- IR pulls **documents** from large text collections (usually the Web) in response to specific keywords or queries. You analyse the **documents**.
- IE pulls **facts** and **structured information** from the content of large text collections. You analyse the **facts**.





# IE for Document Access

- With traditional query engines, getting the facts can be hard and slow
  - Where has the Queen visited in the last year?
  - Which airports are currently closed due to the volcanic ash?
- Which search terms would you use to get these?
- How can you specify you want to see someone's home page?
- IE returns information in a structured way
- IR returns documents containing the relevant information somewhere

# IE as an alternative to IR



- 
- IE returns knowledge at a much deeper level than traditional IR
  - It allows you to specify your query in a more structured way
  - Constructing a database through IE and linking it back to the documents can provide a valuable alternative search tool
  - Even if results are not always accurate, they can be valuable if linked back to the original text



# What is IE used for?

- IE is an enabling technology for many other applications:
  - Text Mining
  - Semantic Annotation
  - Question Answering
  - Opinion Mining
  - Decision Support
  - Rich information retrieval and exploration
  - and so on..





# Two main types of IE systems

## Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- some changes may be hard to accommodate

## Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- require large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

# Named Entity Recognition: the cornerstone of IE



Traditionally, NE is the identification of proper names in texts, and their classification into a set of predefined categories of interest

- Person
- Organisation (companies, government organisations, committees, etc)
- Location (cities, countries, rivers, etc)
- Date and time expressions

Various other types are frequently added, as appropriate to the application, e.g. newspapers, ships, monetary amounts, percentages.



# Why is NE important?

- 
- NE provides a foundation from which to build more complex IE systems
  - Relations between NEs can provide tracking, ontological information and scenario building
  - Tracking (co-reference): “Dr Smith”, “John Smith”, “John”, “he”
  - Ontologies: “Athens, Georgia” vs “Athens, Greece”



# Typical NE pipeline

- 
- Pre-processing (tokenisation, sentence splitting, morphological analysis, POS tagging)
  - Entity finding (gazetteer lookup, NE grammars)
  - Coreference (alias finding, orthographic coreference etc.)
  - Export to database / XML / ontology



# Example of IE

---

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



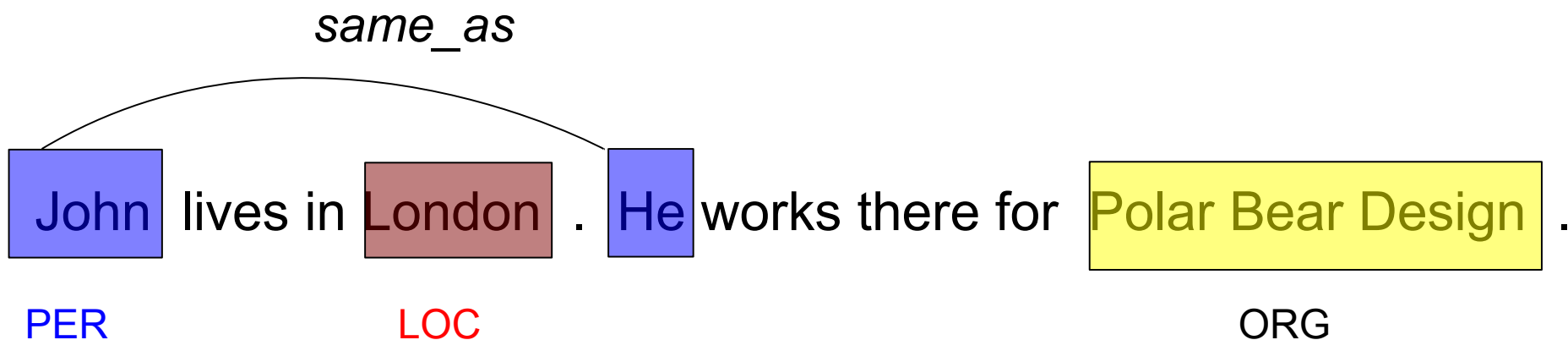
# Basic NE Recognition

---

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

PER LOC ORG

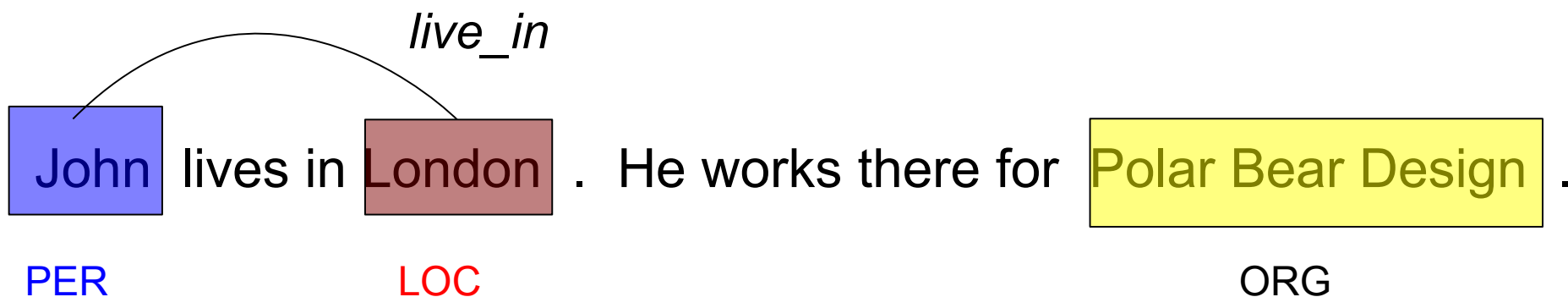
# Co-reference





# Relations

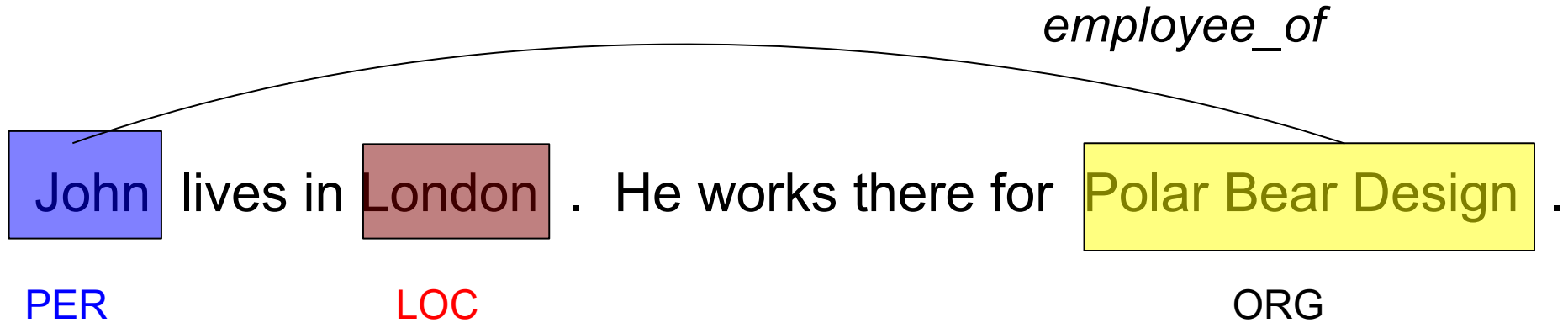
---





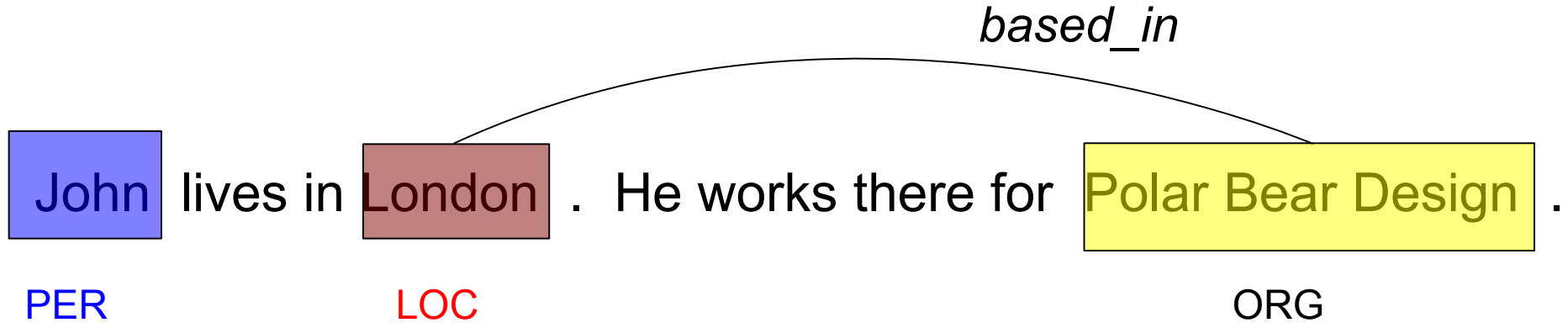


# Relations (2)





# Relations (3)





---

# Examples of IE systems

# HaSIE



- 
- Health and Safety Information Extraction
  - Application developed with GATE, which aims to find out how companies report about health and safety information
  - Answers questions such as:
    - “How many members of staff died or had accidents in the last year?”
    - “Is there anyone responsible for health and safety?”
  - IR returns whole documents



CompanyName

BAA

HSEParagraphs

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

BAA has received a RoSPA gold award

Accidents

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.



# Obstetrics records

- 
- Streamed entity recognition during note taking
    - Interventions, investigations, etc.
  - Based entirely on gazetteers and JAPE
  - Has to cope with terse, ambiguous text and distinguish past events from present
  - Used upstream for decision support and warnings



**GATE**

- Applications
  - pipeline
- Language Resources
- Processing Resources
  - Cleanup
  - Annotation Set Tra
  - IE Transducer
  - Flexible Gazetteer
  - Roots gazetteer

MimeType	text/
currentGravidity	3
day	20
gate.SourceURL	file:/
month	8
shift	12

Messages pipeline Case\_006.htm\_00...

Annotation Sets Annotations List Co-reference Editor Text

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

Type	Set	St
------	-----	----

18 Annotations (0 selected) Select:

Document Editor Initialisation Parameters

- CesareanSectionInPriorDelivery
  - DiastolicBloodPressure
  - DiastolicBloodPressureBefore20W
  - Dinoprostone
  - EstimatedFetalWeight
  - FHREvaluation
  - GBSNeonatalSepsisAfterAPrevious
  - Gravidity
  - HighRiskForAnaphylaxis
  - MagnesiumSulfate
  - MembranesStatus
  - MyastheniaGravis
  - PatientAge
  - PelvicAdequacy
  - PenicillinAllergy
  - PreviousCesareanSectionType
  - SystolicBloodPressure
  - SystolicBloodPressureBefore20We
  - TimeStamp
  - UrineProtein
- New



# Multiflora

---

- IE system in the botanical domain
- Finds information about different plants: size, leaf span, colour etc
- Collates information from different sources: these often refer to plant features in slightly different ways
- Uses shallow linguistic analysis: POS tags and noun and verb phrase chunking
- Important to relate features to the right part of the plant: leaf size rather than plant size, colour of flowers vs colour of leaves etc.



Messages  R\_a\_FNA.txt\_00743

Text Annotations Annotation Sets

Print

7. *Ranunculus acris* Linnaeus, Sp. Pl. 1: 554. 1753

□ Renoncule âcre, bouton d'or

*Ranunculus acris* var. *latisectus* Beck

Stems erect from short caudex or rhizome, never rooting nodally, hispid, strigose, or glabrous, base not bulbous. Roots never tuberous. Basal leaf blades pentagonal in outline, deeply 3-5-parted, 1.8-5.2 X 2.7-9.8 cm, segments 1-2 X -lobed or -parted, ultimate segments narrowly elliptic or oblong to lanceolate, margins toothed or lobulate, apex acute to rounded. Flowers: receptacle glabrous; sepals spreading, 4-6(-9) x 2-5 mm, hispid; petals 5, yellow, 8-11(-17) X 7-13 mm. Heads of achenes globose, 5-7(-10) mm wide; achenes 2-3 X 1.8-2.4 mm, glabrous, margin forming narrow rib 0.1-0.2 mm wide; beak persistent, deltate, usually with tip short or long, straight or curved, subulate, 0.2-1 mm. 2n = 14.

Type	Set	Start ▲	End	Features
PlantFeatures	Default	0	1	{type=number}
Header	Default	0	44	∅
PlantFeatures	Default	38	39	{type=number}
PlantFeatures	Default	103	113	{rule=HeadAdj}
Head	Default	119	124	∅
PlantFeatures	Default	125	130	{rule=HeadAdj}
PlantFeatures	Default	136	141	{rule=AdjHead}
Head	Default	142	148	∅
Head	Default	152	150	∅

Annotations Editor

Features Editor

Default annotations

- Head
- Header
- Lookup
- PlantFeature
- Segment
- SegmentSplit
- Sentence
- SpaceToken
- Split
- Token

Original markups are

- paragraph



# Old Bailey IE

---

- The Old Bailey Proceedings Online makes available a fully searchable, digitised collection of all surviving editions of the Old Bailey Proceedings from 1674 to 1913
- GATE was used to perform IE on the court reports, identifying names of people, places, dates etc.
- ANNIE was customised to only extract full Person names and to take account of old English language used
- More info at <http://www.oldbaileyonline.org/static/Project.jsp>



# Old Bailey IE

Messages | file:/C:/JOB-DataStore/ | 17141209.txt-1.xml\_0004B

Text | Annotations | Annotation Sets | Coreference | Print |

Indictment.

William Mills, of the Parish of St. Sepulchres, was indicted for stealing a dark grey Gelding, value 12 l. out of the Grounds of George More, Esq; on the 5th of October last. It appear'd, That the Horse was lost out of the Prosecutor's Grounds at Newark Trens, and sold by the Prisoner at the G in Smithfield, and he not being able to give an Account how he came by it, was found Guilty of the Indictment.

Laurance Singleton, Mary Singleton, and \_bert, were indicted for breaking the W house of Joseph Wives, and stealing thence 15 Foot Wall out Dlenk 60 Feet of Wainscot and Foot of Deal, on the 29th of September last. It appear'd that \_bert was an Evidence who Swore, he saw him beir them halt a Yard long) and burn them at Singleton's House; which not b Cause for an Indictment, they were acquitted.

Andrew James, (a little Boy) of the Parish of St. Dunstun in the West, wa stealing a Silk Handkerchief, value 2 s. from the P of George Mac, on the 8th instant. It was prov'd that the kerchief taken upon him; whereupon he was found Guilty to the Value of 10 d.

Mary was indicted for Assaulting ) with infection to on the 2nd of November last. It appear'd by

- Default
- Original markups
- Lookup
- Location
- Token
- Sentence

Default annotations

- Date
- FirstPerson
- Foo
- Location
- Lookup
- Organization
- Person
- Sentence
- SpaceToken
- Split
- Temp



# IE in other languages

---

- ANNIE has been adapted to various other languages: some as test cases, some as real IE systems
- More details about this in Track 3 (Advanced IE module)
- Brief introduction to multilingual PRs in GATE later in this tutorial

te

## Applications

arabic not trained

## Language Resources

GATE document\_00095

## Processing Resources

orthomatcher

arabic not trained grammar

arabic gaz

arabic tokeniser

reset

## Data stores

file:/share/nlp.18/diana/ga

GATE document\_00095

file:/share/nlp.18/diana/gatecorpora/arabic/treebank/bbnfiles/test/processed/

## Messages

Text

Annotations

Annotation Sets

Print



نيقوسيا 7-51 (أ ف ر) -0 ابدى نادي فيورتينا الايطالي اهتمامه بضم مهاجم منتخب  
(البرتغال ونادي بنفيكا نونو غوميش (42 عاما

وكانت اندية ارسال الانكليزي وفرغشة التركي وديورتيفو كورونا وريال سوسيداد  
الاسبانيان اعربت عن رغبتها في ضم غوميش الذي قدر بنفيكا قيمة انتقاله بنحو 61 مليون  
دولار .

وقع راديك بايل لاعبي وسط منتخب تشيكا ونادي انتيكو مدريد الاسباني الذي هبط الى الـ  
الدرجة الثانية عقدا انتقل بموجبه الى لنس الفرنسي لمدة 3 سنوات من دون ان تعرف قيمة  
الصفقة.

وكان بايل (72 عاما ) احدا افراد منتخب بلاد ه في كأس الامم الاوروبية الاخيرة لكنه خاض  
031 دقيقة فقط في المباريات الثلاث التي خاضتها تشيكا في البطولة لانها خرجت من الدور  
الاول .

وقع الكرواتي ميلان رايبيتش مهاجم يروجي الايطالي عقدا لمدة سنتين مع فريق الـ  
فرغشة التركي .

. ولغت قيمة انتقال رايبيتش (72 عاما ) نحو 71 مليون دولار

ولعب رايبيتش 31 مباراة دولية مع منتخب بلاد ه وكان في صفوف نادي هايدوك سبليت  
الكرواتي قبل انتقاله الى ايطاليا .

## Default annotations

## Key annotations

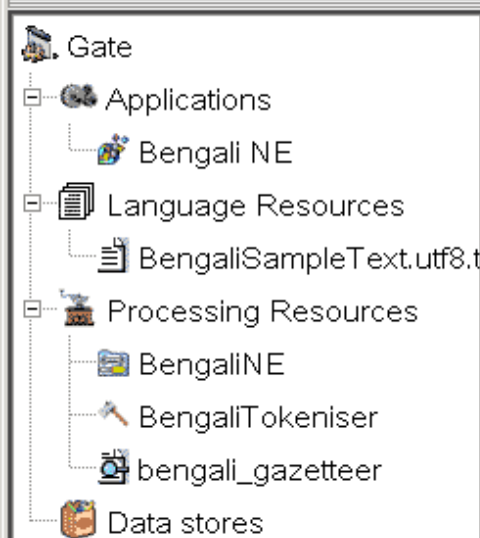
- Cardinal
- Date
- Event
- Gpe
- Gpe\_desc
- Money
- Nationality
- Ordinal
- Org\_desc
- Organization
- Per\_desc
- Person

## Original markups annota

Annotations Editor

Features Editor

Initialisation Parameters



Messages BengaliSampleText.utf8.txt

আমার নাম অনিল রায় । আমি  
লন্ডনকাস্টারে থাকি । আমার বাবা  
লিভারপুলে থাকে ।

আমার বাবার নাম হচ্ছে রাজেশ  
রায় । লন্ডনকাস্টার  
ইউনিভার্সিটি আমার পদার যাওয়া  
। আমার বাবা কংকো কংগো  
কম্পনিতে কাজ করে ।

My name is Anil Roy. I live in Lancaster. My father lives in  
Liverpool. My  
father's name is Rajesh Roy. Lancaster University is my place of

Type	Set	Start ▲	End	Features
Person	Default	10	18	{kind=fullname}
Location	Default	27	38	{kind=city, rule=City}
Location	Default	59	67	{kind=city, rule=City}
Person	Default	101	112	{kind=fullname}
Organisation	Default	115	141	{}
Organisation	Default	173	182	{}

Annotations Features

Default annotations

- DEFAULT\_TOKEN
- Location
- Lookup
- Organisation
- Person
- SpaceToken
- Token



---

# **ANNIE: A Nearly New Information Extraction system**



# About this tutorial

---

- As before, this tutorial will be a hands on session with some explanation as you go.
- We will use a corpus of news texts in the file [module-2-hands-on.zip](#). Unzip this file if it isn't already.
- Things for you to try yourself are in **red**.
  - There will be instructions for you to follow for each step
  - Each step will be demonstrated
  - Correct answers will be shown before moving on
- **Start GATE on your computer now (if you haven't already)**





# Extra exercises

---

- We need to pace the exercises for everyone.
- If it is too slow for you please feel free to skip through the exercises at your own pace.
- If you get a long way ahead, there are extra exercises at the end of these slides
  - You may not be able to do these extra exercises until you have finished the main tutorial exercises
  - You do not need to do this extra material to complete the tutorial. It is not a prerequisite for the rest of the course.



# Extra exercises

---

- Note that instructions for the extra exercises are briefer than for the rest of the tutorial – they assume you now have the basics of GATE
- The extra exercises are:
  - Comparing different sentence splitters
  - Further evaluation exercises
  - Using the QA tools to compare three IE systems
    - ANNIE
    - LingPipe
    - OpenNLP
  - Demonstration of an ontology based gazetteer



# Nearly New Information Extraction

---

- ANNIE is a ready made collection of PRs that performs IE on unstructured text.
- For those who grew up in the UK, you can think of it as a Blue Peter-style “here's one we made earlier”.
- ANNIE is “nearly new” because
  - It was based on an existing IE system, LaSIE
  - We rebuilt LaSIE because we decided that people are better than dogs at IE
  - Being 10 years old, it's not really new any more

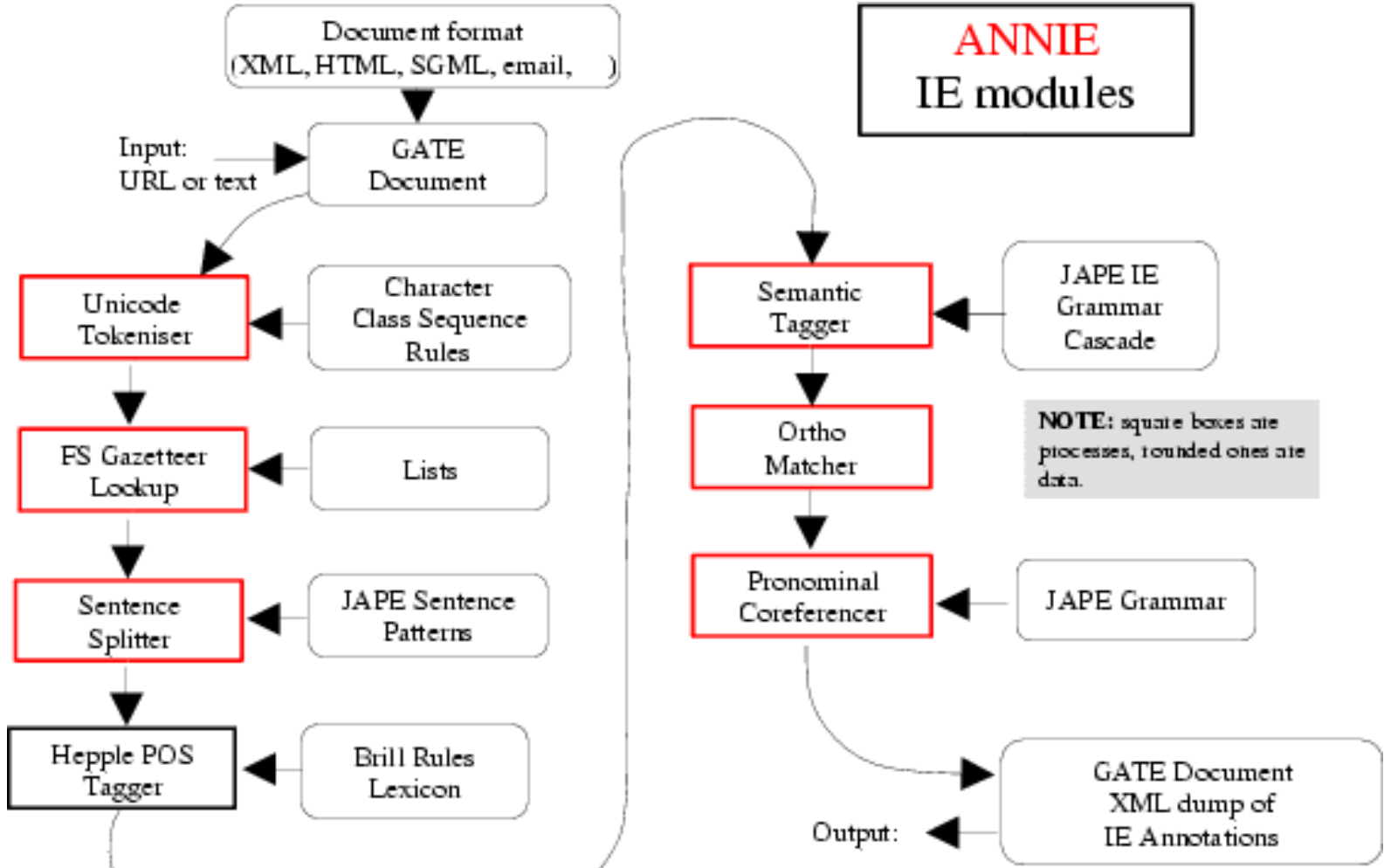


# What's in ANNIE?

- 
- The ANNIE application contains a set of core PRs:
    - Tokeniser
    - Sentence Splitter
    - POS tagger
    - Gazetteers
    - Named entity tagger (JAPE transducer)
    - Orthomatcher (orthographic coreference)
  - There are also other PRs available in the ANNIE plugin, which are not used in the default application, but can be added if necessary
    - NP and VP chunker




# Core ANNIE components





# Loading and running ANNIE

- Because ANNIE is a ready-made application, we can just load it directly from the menu
- Click the  icon from the top GATE menu OR  
File → Ready Made Applications → ANNIE → ANNIE OR  
right-click Applications → Ready Made Applications → ANNIE → ANNIE
- Select “with defaults” if necessary
- Load the hands-on corpus from the “news-texts” directory in the zip file
- Run ANNIE and inspect the annotations
- You should see a mixture of Named Entity annotations (Person, Location etc) and some other linguistic annotations (Token, Sentence etc)



# Let's look at the PRs

- 
- Each PR in the ANNIE pipeline creates some new annotations, or modifies existing ones
  - Document Reset → removes annotations
  - Tokeniser → Token annotations
  - Gazetteer → Lookup annotations
  - Sentence Splitter → Sentence, Split annotations
  - POS tagger → adds category features to Token annotations
  - NE transducer → Date, Person, Location, Organisation, Money, Percent annotations
  - Orthomatcher → adds match features to NE annotations



# Document Reset

---

- This PR should go at the beginning of (almost) every application you create
- It removes annotations created previously, to prevent duplication if you run an application more than once
- It does not remove the Original Markups set, by default
- You can configure it to keep any other annotation sets you want, or to remove particular annotation types only





# Document Reset Parameters

Loaded Processing resources

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

Selected Processing resources

!	Name	Type
<input checked="" type="checkbox"/>	Document Reset PR_00016	Docur

Run "Document Reset PR\_00016"?

Yes  No  If value of feature  is

Corpus: <none>

Runtime Parameters for the "Document Reset PR\_00016" Document Reset PR:

Name	Type	Required	Value
<input type="checkbox"/> annotationTypes	ArrayList		[]
<input type="checkbox"/> keepOriginalMarkupsAS	Boolean		true
<input type="checkbox"/> setsToKeep	ArrayList		[Key ]

Run this Application

Specify any specific annotations to remove. By default, remove all.

Keep Original Markups set

Keep Key set



---

# Tokenisation and sentence splitting



# Tokeniser

- Tokenisation based on Unicode classes
- Declarative token specification language
- Produces Token and SpaceToken annotations with features orthography and kind
- Length and string features are also produced
- Rule for a lowercase word with initial uppercase letter

```
"UPPERCASE_LETTER" LOWERCASE_LETTER"* >  
  Token; orthography=upperInitial; kind=word
```



# Document with Tokens

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

Union Appeals For Talks To End BA Strike

Skip to navigation | Skip to content |  
Home | Contact Us | News Search;  
HubPage  
Airwise News  
Airport Guide  
Airwise Travel  
Search  
Union Appeals For Talks To End BA Strike  
March 22, 2010

Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers.

Type	Features
Token	{ category=NNP, kind=word, length=5, orth=upperInitial, string=Union}
Token	{ category=NNPS, kind=word, length=7, orth=upperInitial, string=Appeals}
Token	{ category=IN, kind=word, length=3, orth=upperInitial, string=For}
Token	{ category=NNS, kind=word, length=5, orth=upperInitial, string=Talks}
Token	{ category=TO, kind=word, length=2, orth=upperInitial, string=To}

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- Original markups



# ANNIE English Tokeniser

---

- The English Tokeniser is a slightly enhanced version of the Unicode tokeniser
- It comprises an additional JAPE transducer which adapts the generic tokeniser output for the POS tagger requirements
- It converts constructs involving apostrophes into more sensible combinations
  - don't → do + n't
  - you've → you + 've



# Looking at Tokens

---

- Tidy up GATE by removing all resources and applications (or just restart GATE)
- Load the news text hands-on corpus
- Create a new application (corpus pipeline)
- Load a Document Reset and an ANNIE English Tokeniser
- Add them (in that order) to the application and run on the corpus
- View the Token and SpaceToken annotations
- What different values of the “kind” feature do you see?



# Sentence Splitter

---

- The default splitter finds sentences based on Tokens
- Creates Sentence annotations and Split annotations on the sentence delimiters
- Uses a gazetteer of abbreviations etc. and a set of JAPE grammars which find sentence delimiters and then annotate sentences and splits
- Load an ANNIE Sentence Splitter PR and add it to your application (at the end)
- Run the application and view the results



# Document with Sentences

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

the opposition conservatives, ahead in opinion polls, have been turning up the pressure on Labour over its links to Unite, saying the government had failed to take action quickly enough because it did not want to alienate its financial backers.

"We deplore the strike, and the prime minister and the transport secretary have said that absolutely clearly," Foreign Secretary David Miliband told Sky News.

"The way to resolve these disputes is through negotiation, it is damaging for the company, it is damaging for the crews and it is damaging for the country."

The dispute arose because BA, which has 12,000 cabin crew, wants to save an annual GBP£62.5 million pounds (USD\$95 million) to help cope with a fall in demand, volatile fuel prices and increased competition from low-cost carriers.

A spokesman said there was no estimate yet as to how much the industrial action would cost the company.

Type	Features
Sentence {}	
Sentence {}	
Sentence {}	
Sentence {}	
Sentence {}	

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- ▶ Original markups





# Sentence splitter variants

---

- An alternate set of rules can be loaded with the regular sentence splitter
- To do this, reload the sentence splitter using “main-single-nl.jape” instead of “main.jape” as the value of the grammar parameter
- The main difference is the way it handles new lines
- In some cases, you might want a new line to signal a new sentence, e.g. addresses
- In other cases, you might not, e.g. in emails that have been split by the email program
- A regular expression Java-based splitter is also available, called RegEx Sentence Splitter, which is sometimes faster
- This handles new lines in the same way as the default sentence splitter
- See “Further Exercises” to experiment with splitter variants



---

# Shallow lexico-syntactic features



# POS tagger

- 
- ANNIE POS tagger is a Java implementation of Brill's transformation based tagger
  - Previously known as **Hepple Tagger** (you may find references to this and to **heptag**)
  - Trained on WSJ, uses Penn Treebank tagset
  - Default ruleset and lexicon can be modified manually (with a little deciphering)
  - Adds category feature to Token annotations
  - Requires Tokeniser and Sentence Splitter to be run first



# Morphological analyser

---

- Not an integral part of ANNIE, but can be found in the Tools plugin as an “added extra”
- Flex based rules: can be modified by the user (instructions in the User Guide)
- Generates “root” feature on Token annotations
- Requires Tokeniser to be run first
- Requires POS tagger to be run first if the considerPOSTag parameter is set to true

# Shallow lexico-syntactic features



- 
- Add an ANNIE POS Tagger to your app
  - Add a GATE Morphological Analyser after the POS Tagger
  - If this PR is not available, load the Tools plugin first
  - Re-run your application
  - Examine the features of the Token annotations
    - New features of category and root have been added



---

# Gazetteers



# Gazetteers

- Gazetteers are plain text files containing lists of names (e.g rivers, cities, people, ...)
- The lists are compiled into Finite State Machines
- Each gazetteer has an index file listing all the lists, plus features of each list (majorType, minorType and language)
- Lists can be modified either internally using the Gazetteer Editor, or externally in your favourite editor (note that the new Gazetteer editor replaces the old GAZE editor you may have seen previously)
- Gazetteers generate Lookup annotations with relevant features corresponding to the list matched
- Lookup annotations are used primarily by the NE transducer
- Various different kinds of gazetteer are available: first we'll look at the default ANNIE gazetteer

# Running the ANNIE Gazeteer

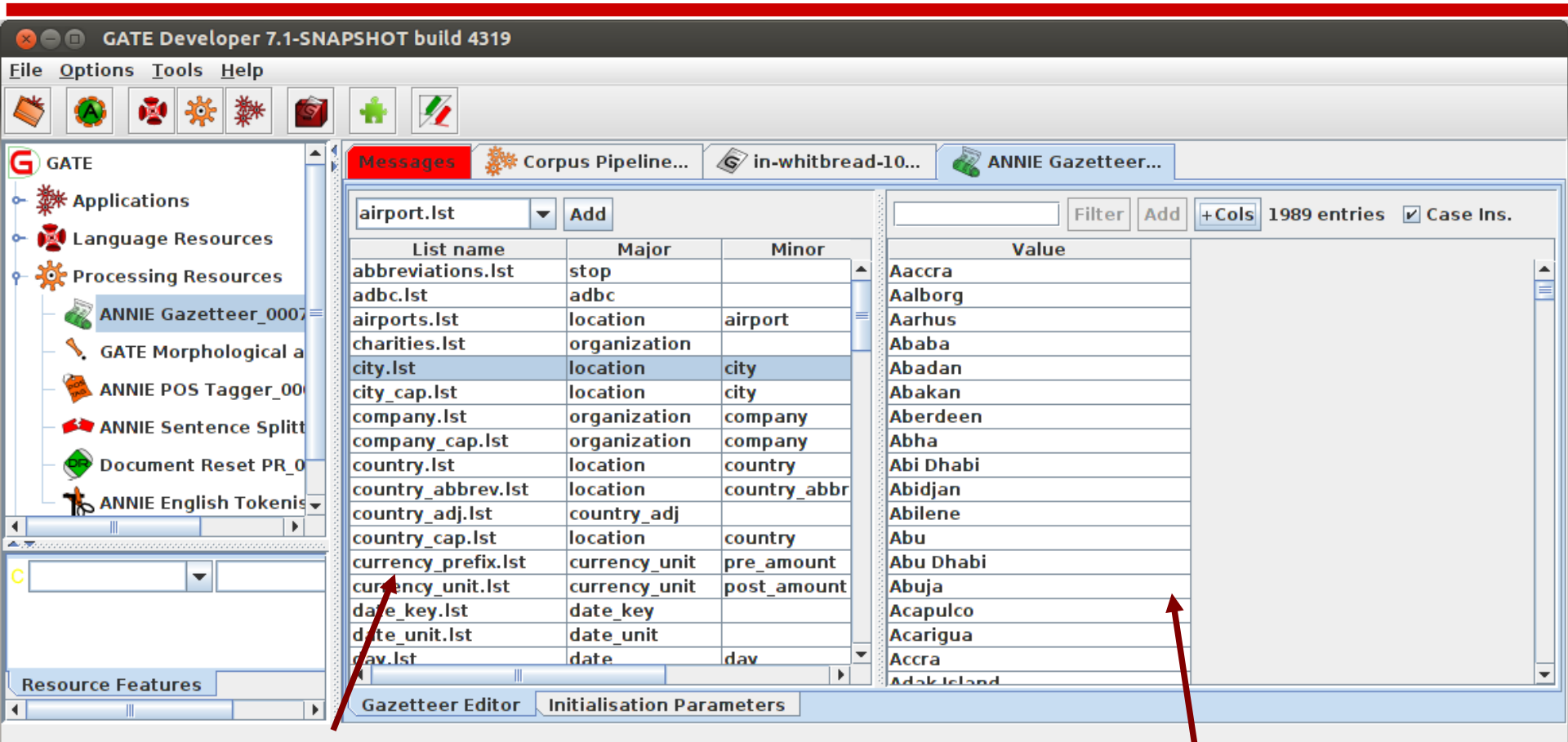


- 
- Various different kinds of gazetteer are available: first we'll look at the default ANNIE gazetteer
  - Load the ANNIE Gazetteer PR and double click on it in the resource pane to open and see the gazetteers
  - Add it to the end of your pipeline
  - Re-run the pipeline
  - Look for “Lookup” annotations and examine their features





# Gazetteer editor



definition file entries

entries for selected list



# ANNIE gazetteer - contents

---

- Double click on the ANNIE Gazetteer PR (under Processing Resources in the left hand pane) to open it
- Select “Gazetteer Editor” from the bottom tab
- In the left hand pane (linear definition) you see the index file containing all the lists
- In the right hand pane you see the contents of the list selected in the left hand pane
- Each entry can be edited by clicking in the box and typing
- New entries can be added by typing in the “New list” or “New entry” box respectively

# Modifying the definition file

add a new list

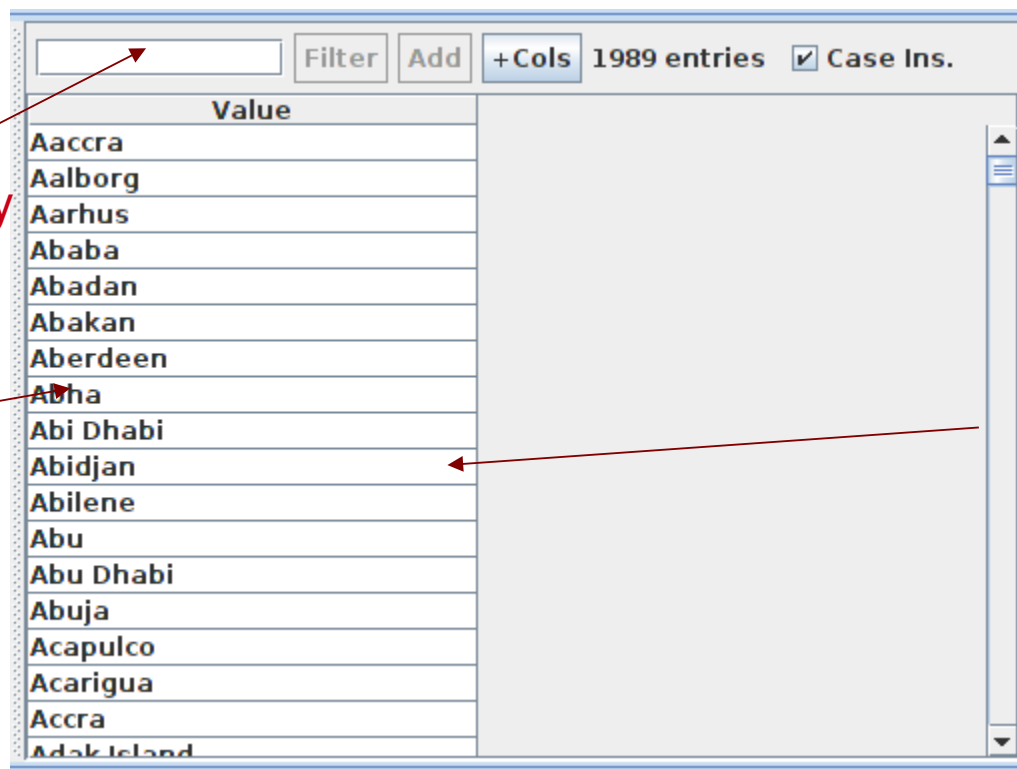
edit an existing list name by typing here

delete a list by right clicking on an entry and selecting Delete

List name	Major	Minor
abbreviations.lst	stop	
adbc.lst	adbc	
airports.lst	location	airport
charities.lst	organization	
city.lst	location	city
city_cap.lst	location	city
company.lst	organization	company
company_cap.lst	organization	company
country.lst	location	country
country_abbrev.lst	location	country_abbrev
country_adj.lst	country_adj	
country_cap.lst	location	country
currency_prefix.lst	currency_unit	pre_amount
currency_unit.lst	currency_unit	post_amount
date_key.lst	date_key	
date_unit.lst	date_unit	
day.lst	date	day

edit the major and minor Types by typing here

# Modifying a list



add a new entry  
by typing here

edit an  
existing entry  
by typing here

Delete an entry by  
right clicking and  
selecting "Delete"



# Editing gazetteer lists

- The ANNIE gazetteer has about 60,000 entries arranged in 80 lists
- Each list reflects a certain category, e.g. airports, cities, first names etc.
- List entries might be entities or parts of entities, or they may contain contextual information (e.g. job titles often indicate people)
- **Click on any list to see the entries**
- Note that some lists are not very complete!
- **Try adding, deleting and editing existing lists, or the list definition file**
- **To save an edited gazetteer, right click on the gazetteer name in the tabs at the top or in the resources pane on the right, and select “Save and Reinitialise” before running the gazetteer again.**
- **Try adding a word from a document you have loaded (that is not currently recognised as a Lookup) into the gazetteer, re-run the gazetteer and check the results.**

# Editing gazetteers outside GATE



- 
- You can also edit both the definition file and the lists outside GATE, in your favourite text editor
  - If you choose this option, you will need to reinitialise the gazetteer in GATE before running it again
  - To reinitialise any PR, right click on its name in the Resources pane and select “Reinitialise”



# List attributes

---

- When something in the text matches a gazetteer entry, a Lookup annotation is created, with various features and values
- The ANNIE gazetteer has the following default feature types: majorType, minorType, language
- These features are used as a kind of classification of the lists: in the definition file features are separated by “:”
- For example, the “city” list has a majorType “location” and minorType “city”, while the “country” list has “location” and “country” as its types
- Later, in the JAPE grammars, we can refer to all Lookups of type location, or we can be more specific and refer just to those of type “city” or type “country”



# Ontologies in IE

---

- A typical way to use an ontology in IE is to create a gazetteer from names and labels in the ontology, and use this to annotate entities with IDs (URIs) from the ontology
- GATE includes several tools to help with this, including a basic ontology viewer and editor, several ontology backed gazetteers, and the ability to refer to ontology classes in grammars
- This is covered in detail in Track 3, Module 9
- The extra exercises includes an example for you to try, a simple demo application that creates a gazetteer from a SPARQL endpoint, adds entity annotations, and then adds further information to the entities, from the ontology





---

# NE transducers



# NE transducer

- Gazetteers can be used to find terms that suggest entities
- However, the entries can often be ambiguous
  - “May Jones” vs “May 2010” vs “May I be excused?”
  - “Mr Parkinson” vs “Parkinson's Disease”
  - “General Motors” vs. “General Smith”
- Handcrafted grammars are used to define patterns over the Lookups and other annotations
- These patterns can help disambiguate, and they can combine different annotations, e.g. Dates can be comprised of day + number + month
- NE transducer consists of a number of grammars written in the JAPE language
- Module 3 tomorrow will be devoted to JAPE



# ANNIE NE Transducer

---

- Load an ANNIE NE Transducer PR
- Add it to the end of the application
- Run the application
- Look at the annotations
- You should see some new annotations such as Person, Location, Date etc.
- These will have features showing more specific information (eg what kind of location it is) and the rules that were fired (for ease of debugging)



# ANNIE NE Transducer

---

- Load an ANNIE NE Transducer PR
- Add it to the end of the application
- Run the application
- Look at the annotations
- You should see some new annotations such as Person, Location, Date etc.
- These will have features showing more specific information (eg what kind of location it is) and the rules that were fired (for ease of debugging)



---

# Co-reference



# Using co-reference

- 
- Different expressions may refer to the same entity
  - Orthographic co-reference module (orthomatcher) matches proper names and their variants in a document
  - [Mr Smith] and [John Smith] will be matched as the same person
  - [International Business Machines Ltd.] will match [IBM]



# Orthomatcher PR

- Performs co-reference resolution based on orthographical information of entities
- Produces a list of annotation IDs that form a co-reference “chain”
- List of such lists stored as a document feature named “MatchesAnnots”
- Improves results by assigning entity type to previously unclassified names, based on relations with classified entities
- May not reclassify already classified entities
- Classification of unknown entities very useful for surnames which match a full name, or abbreviations, e.g. “Bonfield” <Unknown> will match “Sir Peter Bonfield” <Person>
- A pronominal PR is also available



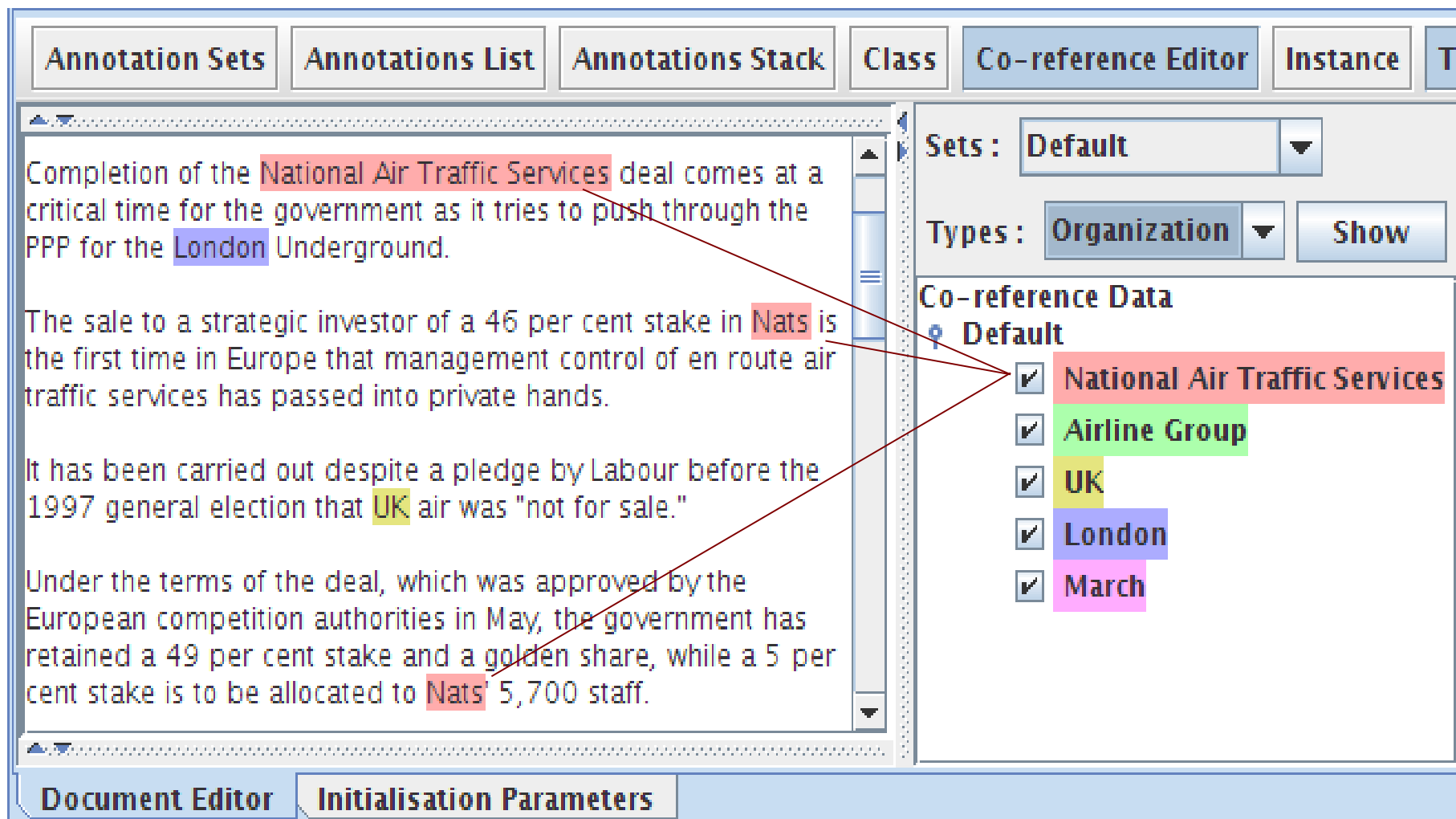
# Looking at co-reference

---

- Add a new PR: ANNIE OrthoMatcher
- Add it to the end of the application
- Run the application
- In a document view, open the co-reference editor by clicking the button above the text
- All the documents in the corpus should have some co-reference, but some may have more than others



# Coreference editor

The screenshot shows the GATE Coreference Editor interface. At the top, there is a navigation bar with buttons for "Annotation Sets", "Annotations List", "Annotations Stack", "Class", "Co-reference Editor" (which is highlighted), "Instance", and "T...". Below this, the main window is divided into two panes. The left pane displays a document with several paragraphs of text. The right pane shows the "Co-reference Editor" settings. The "Sets" dropdown is set to "Default". The "Types" dropdown is set to "Organization", and a "Show" button is visible. Below these, the "Co-reference Data" section shows a list of entities with checkboxes and colored highlights. The entities are: "National Air Traffic Services" (red highlight), "Airline Group" (green highlight), "UK" (yellow highlight), "London" (purple highlight), and "March" (pink highlight). Red lines connect the highlighted text in the document to the corresponding entries in the Co-reference Data list. At the bottom of the interface, there are buttons for "Document Editor" and "Initialisation Parameters".

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance T...

Sets: Default

Types: Organization Show

Co-reference Data

Default

- National Air Traffic Services
- Airline Group
- UK
- London
- March

Document Editor Initialisation Parameters

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.

# Using the co-reference editor



- 
- Select the annotation set you wish to view (Default)
  - A list of all the co-reference chains that are based on annotations in the currently selected set is displayed
  - Select an item in the list to highlight all the member annotations of that chain in the text (you can select more than one at once)
  - Hovering over a highlighted annotation in the text enables you to Delete an item from the co-reference chain
  - Try it!

# Using the co-reference editor



- 
- Deselect all items in the coreference list (right hand pane), then select a type from the “Type” combo box (e.g. “Person”) and click “Show” to view all coreferences of a particular annotation type (note that some types may not have coreferences)
  - Hovering over a highlighted annotation in the text enables you to add a coreference between this annotation and one of the coreference chains listed in the right hand pane
  - Try it!



---

# Modifying ANNIE




# Modifying ANNIE

---

- Typically any new application you want to create will use some or all of the core components from ANNIE
- The tokeniser, sentence splitter and orthomatcher are basically language, domain and application-independent
- The POS tagger is language dependent but domain and application-independent
- You may also require additional PRs (either existing or new ones – e.g. morphological analyser)
- The gazetteer lists and JAPE grammars may act as a starting point but will almost certainly need to be modified



# ANNIE without defaults

- This option loads all the ANNIE PRs, but enables you to change the location of any of them
- It's useful if you want to use ANNIE but you want to change some of the PRs slightly or replace them with your own modified versions
- Restart GATE or remove all PRs and applications, to tidy up a little
- In your file browser or on the command line, look for `plugins/ANNIE/resources/gazetteer` in your GATE home directory
- Copy the whole gazetteer directory to a new location on your computer and make some changes to the lists and/or to the index in a text editor
- Load ANNIE from  but select "Without defaults"
- For each PR, select the default option, except for the gazetteer, where you should select your saved gazetteer index file (`lists.def`)



---

# Multilingual IE



# Language plugins

- Language plugins contain language-specific PRs, with varying degrees of sophistication and functions for:
  - Arabic
  - Cebuano
  - Chinese
  - Hindi
  - Romanian
- There are also various applications and PRs available for French, German and Italian
- These do not have their own plugins as they do not provide new kinds of PR
- Applications and individual PRs for these are found in gate/plugins directory: load them as any other PR
- More details of language plugins in user guide



# Building a language-specific application



- The following PRs are largely language-independent:
  - Unicode tokeniser
  - Sentence splitter
  - Gazetteer PR (but do localise the lists!)
  - Orthomatcher (depending on the nature of the language)
- Other PRs will need to be adapted (e.g. JAPE transducer) or replaced with a language-specific version (e.g. POS tagger)
- This topic is covered in more detail in Track 3 (Advanced IE module)



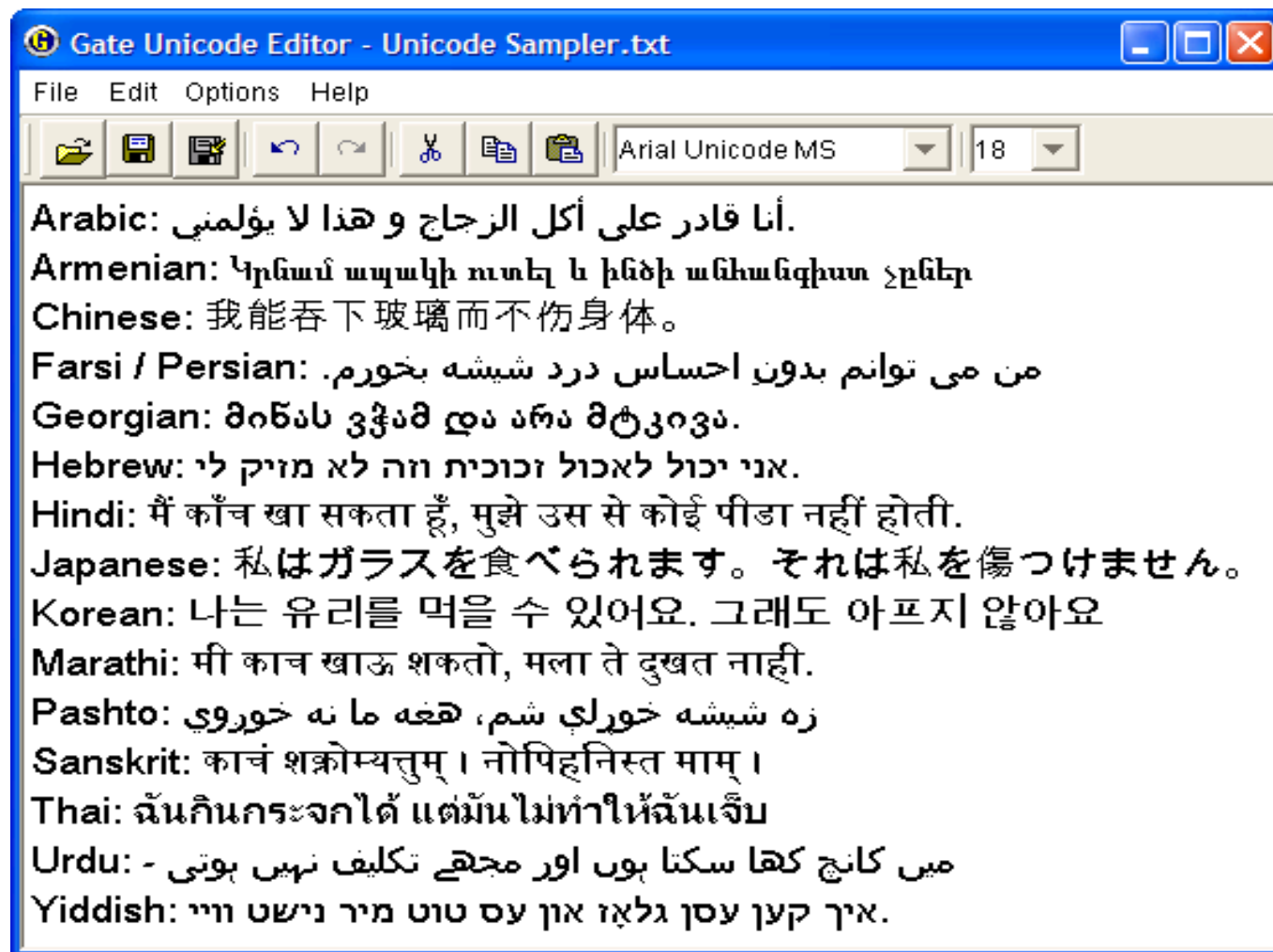
# Useful Multilingual PRs

---

- Stemmer plugin
  - Consists of a set of stemmer PRs for: Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish
  - Requires Tokeniser first (Unicode one is best)
  - Language is init-time param, which is one of the above in lower case
- TreeTagger
  - a language-independent POS tagger which supports English, French, German and Spanish in GATE

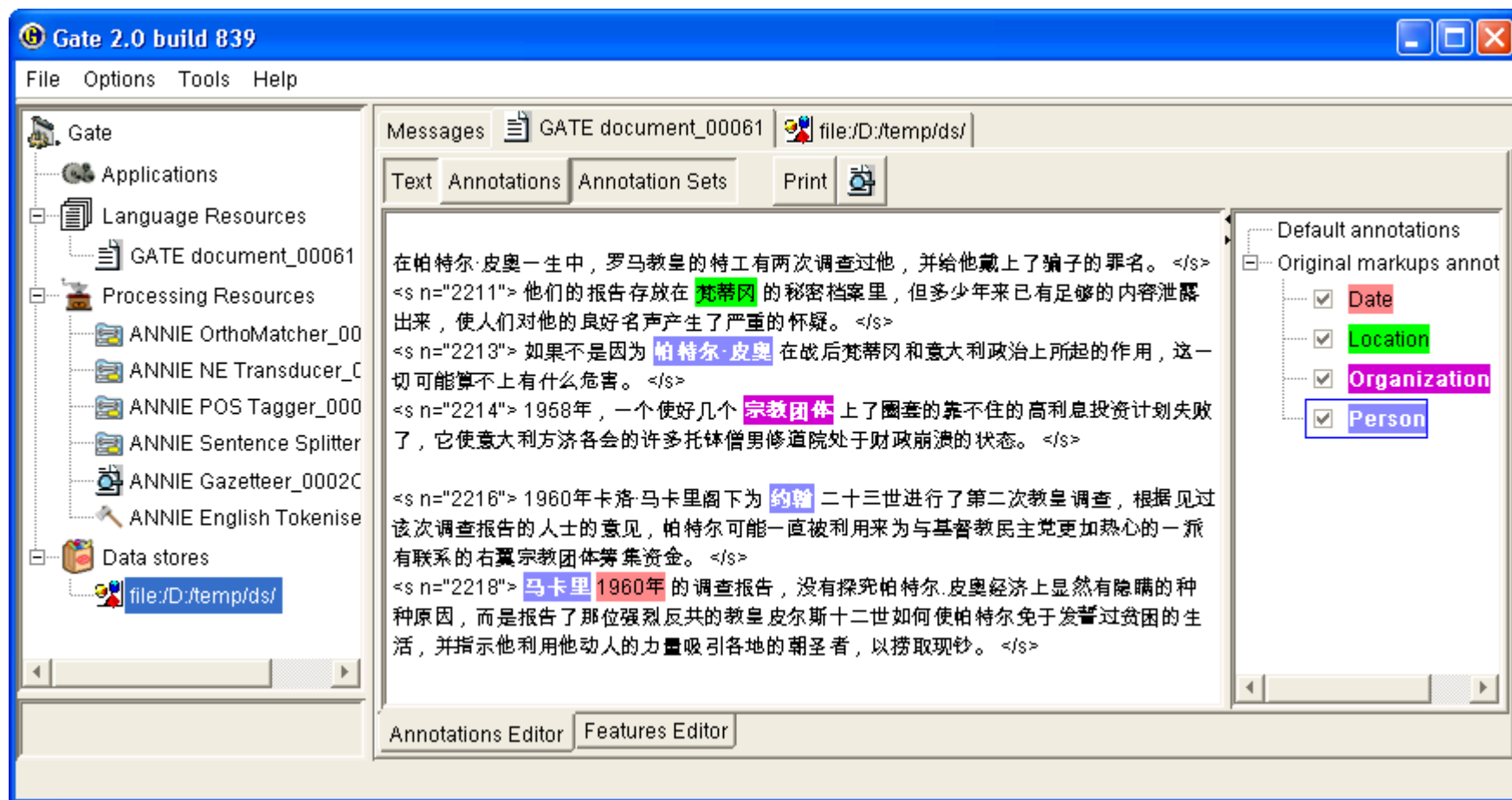
# Displaying multilingual data

GATE uses standard (and imperfect) Java rendering engine for displaying text in multiple languages.



# Displaying multilingual data

All visualisation and editing tools use the same facilities



Gate 2.0 build 839

File Options Tools Help

Messages GATE document\_00061 file:/D:/temp/ds/

Text Annotations Annotation Sets Print

在伯特尔·皮奥一生中，罗马教皇的特工有两次调查过他，并给他戴上了骗子的罪名。 </s>  
 <s n="2211"> 他们的报告存放在 梵蒂冈 的秘密档案里，但多少年来已有足够的内容泄露出来，使人们对他的良好名声产生了严重的怀疑。 </s>  
 <s n="2213"> 如果不是因为 伯特尔·皮奥 在战后梵蒂冈和意大利政治上所起的作用，这一切可能算不上有什么危害。 </s>  
 <s n="2214"> 1958年，一个使好几个 宗教团体 上了圈套的靠不住的高利息投资计划失败了，它使意大利方济各会的许多托钵僧男修道院处于财政崩溃的状态。 </s>  
 <s n="2216"> 1960年卡洛·马卡里阁下为 约翰 二十三世进行了第二次教皇调查，根据见过该次调查报告的人士的意见，伯特尔可能一直被利用来为与基督教民主党更加热心的一派有联系的右翼宗教团体筹集资金。 </s>  
 <s n="2218"> 马卡里 1960年 的调查报告，没有探究伯特尔·皮奥经济上显然有隐瞒的种种原因，而是报告了那位强烈反共的教皇皮尔斯十二世如何使伯特尔免于发誓过贫困的生活，并指示他利用他动人的力量吸引各地的朝圣者，以捞取现钞。 </s>

Default annotations

- Original markups annot
  - Date
  - Location
  - Organization
  - Person

Annotations Editor Features Editor



---

# Annotation and Evaluation



# Topics covered

---

- Defining annotation guidelines
- Recap on manual annotation using the GATE GUI
- Using the GATE evaluation tools

# Before you start annotating

---

- You need to think about annotation guidelines
- You need to consider what you want to annotate and then to define it appropriately
- With multiple annotators it's essential to have a clear set of guidelines for them to follow
- Consistency of annotation is really important for a proper evaluation



# Annotation Guidelines

- 
- People need clear definition of what to annotate in the documents, with examples
  - Typically written as a guidelines document
  - Piloted first with few annotators, improved, then “real” annotation starts, when all annotators are trained
  - Annotation tools may require the definition of a formal DTD (e.g. XML schema)
    - What annotation types are allowed
    - What are their attributes/features and their values
    - Optional vs obligatory; default values





# Annotation Editor

File Options Tools Help

file:/home/dian... S01121~O\_0024A S0FNTC~D\_00250

Messages rename-FAO-anno...

Annotation Sets Annotations List Co-reference Editor OAT Text

This species reaches a maximum size of 445 cm total length and about 540 kg weight. The size range of fish taken by the commercial swordfish longliners is 120 to 190 cm body length in the northwestern Pacific; the average weight in the Mediterranean Sea ranges from 115 to 160 kg. Usually females are larger than males, and most swordfish over 140 kg are females. Adults grow over 230 kg (rarely) in the Mediterranean, up to 320 kg in the western Atlantic, and up to 537 kg in the southeast. The all-tackle-angling record for this species is a 536 kg fish caught off Iquique, Chile in 1953. There is little biological minimum size and age and some of the

- Key
- Location
- Original markups

Type	Set	Start	End	Id	Features
Location	Key	3067	3084	850	{kind=water}

Location

kind water

Open Search & Annotate tool

1 Annotations (1 selected) Select: [ ] New

Document Editor Initialisation Parameters

Views built!



# Annotation Recap

---

- Adding annotation sets
- Adding annotations
- Resizing them (changing boundaries)
- Deleting
- Changing highlighting colour
- Setting features and their values
- Using the co-reference editor

# Evaluation



“We didn’t underperform. You overexpected.”

# Performance Evaluation



2 main requirements:

- **Evaluation metric:** mathematically defines how to measure the system's performance against human-annotated gold standard
- **Scoring program:** implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of annotation

# Evaluation exercises: preparation

- Restart GATE, or close all documents and PRs to tidy up
- Load the hands on corpus
- Take a look at the annotations.
- There is a set called “Key”. This is a set of annotations against which we want to evaluate ANNIE. In practice, they could be manual annotations, or annotations from another application.
- Load the ANNIE system with defaults
- **Important:** Change the runtime parameters for the Document Reset PR, adding “Key” to the setsToKeep parameter. This stops the application deleting our Key annotations when we run it.
- Run ANNIE: You should have annotations in the Default set from ANNIE, and in the Key set, against which we can compare them.



# AnnotationDiff

- 
- Graphical comparison of 2 sets of annotations
  - Visual diff representation, like tkdiff
  - Compares one document at a time, one annotation type at a time

# Annotations are like squirrels...



Annotation Diff helps with “spot the difference”



# Annotation Diff Exercise

- 
- Open the document “ft-airlines-27-jul-2001.xml”
  - Open the AnnotationDiff (Tools → Annotation Diff or click the icon)
  - For the Key set (containing the manual annotations) select **Key** annotation set
  - For the Response set (containing annotations from ANNIE) select **Default** annotation set
  - Select the **Organization** annotation
  - Click on “Compare”
  - Scroll down the list, to see correct, partially correct, missing and spurious annotations





# Annotation Diff

**Annotation Diff Tool**

Key doc:  Key set:  Type:  Weight:

Resp. doc:  Resp. set:  Features:  all  some  none

Start	End	Key	Features	=?	Start	End	
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

Correct: 19      Recall Precision F-measure

Partially correct: 7      Strict: 0.68 0.68 0.68

Missing: 2      Lenient: 0.93 0.93 0.93

False positives: 2      Average: 0.80 0.80 0.80

10 documents loaded



# A Word about Terminology

---

- Different communities use different terms when talking about evaluation, because the tasks are a bit different.
- The IE community usually talks about “correct”, “spurious” and “missing”
- The IR community usually talks about “true positives”, “false positives” and “negatives”. They also talk about “false negatives”, but you can ignore those.
- Some terminologies assume that one set of annotations is correct (“gold standard”)
- Other terminologies do not assume one annotation set is correct
- When measuring inter-annotator agreement, there is no reason to assume one annotator is more correct than the other



# Terminology Comparison

---

## IE metrics

Correct

Missing

Spurious

Partially correct

## IR metrics

True Positive

False Negative

False Positive

True negative

## Inter-annotator agreement

Match

Only A (or B)

Only B (or A)

Overlap



# Measuring success

---

- In IE, we classify the annotations produced in one of 4 ways:
- **Correct** = things annotated correctly  
e.g. annotating “Hamish Cunningham” as a Person
- **Missing** = things not annotated that should have been  
e.g. not annotating “Sheffield” as a Location
- **Spurious** = things annotated wrongly  
e.g. annotating “Hamish Cunningham” as a Location
- **Partially correct** = the annotation type is correct, but the span is wrong  
e.g, annotating just “Cunningham” as a Person (too short) or  
annotating “Unfortunately Hamish Cunningham” as a Person (too long)

# Finding Precision, Recall and F-measure



Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: 1.0

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features: all some none 1.0

Start	End	Key	Features	=?	Start	End	
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

10 documents loaded

Correct:	19	Recall	Precision	F-measure	
Partially correct:	7	Strict:	0.68	0.68	0.68
Missing:	2	Lenient:	0.93	0.93	0.93
False positives:	2	Average:	0.80	0.80	0.80

Statistics Adjudication

← scores displayed



# Precision

---

- How many of the entities your application found were correct?
- Sometimes precision is called **accuracy**

$$\textit{Precision} = \frac{\textit{Correct}}{\textit{Correct} + \textit{Spurious}}$$



# Recall

- How many of the entities that exist did your application find?
- Sometimes recall is called **coverage**

$$\text{Recall} = \frac{\text{Correct}}{\text{Correct} + \text{Missing}}$$



# F - Measure

---

- Precision and recall tend to trade off against one another
  - If you specify your rules precisely to improve precision, you may get a lower recall
- If you make your rules very general, you get good recall, but low precision
- This makes it difficult to compare applications, or to check whether a change has improved or worsened the results overall
- F-measure combines precision and recall into one measure



# F-Measure

- Also known as the “harmonic mean”
- Usually, precision and recall are equally weighted
- This is known as F1
- To use F1, set the value of the F-measure weight to 1
- This is the default setting

$$F = 2 \cdot \left( \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} \right)$$



# Annotation Diff defaults to F1

Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: 1.0

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features: all some none 1.0 Compare

Start	End	Key	Features	=/?	Start	End	
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Arc
				?-	2386	2395	Hampshire

10 documents loaded

Correct:	19	Recall	Precision	F-measure
Partially correct:	7	Strict:	0.68	0.68
Missing:	2	Lenient:	0.93	0.93
False positives:	2	Average:	0.80	0.80

Statistics Adjudication

F-measure weight set to 1



# Statistics can mean what you want them to....

- How we want to measure partially correct annotations may differ, depending on our goal
- In GATE, there are 3 different ways to measure them
- The most usual way is to consider them to be “half right”
- Average: Strict and lenient scores are averaged (this is the same as counting a half weight for every partially correct annotation)
- Strict: Only perfectly matching annotations are counted as correct
- Lenient: Partially matching annotations are counted as correct. This makes your scores look better :-)



# Strict, Lenient and Average

**Annotation Diff Tool**

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: 1.0

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features:  all  some  none

Start	End	Key	Features	=?	Start	End	
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

10 documents loaded

Correct:	19	Recall	Precision	F-measure
Partially correct:	7	Strict:	0.68	0.68
Missing:	2	Lenient:	0.93	0.93
False positives:	2	Average:	0.80	0.80

Statistics | Adjudication

# Comparing the individual annotations

---



- In the AnnotationDiff, colour codes indicate whether the annotation pair shown are correct, partially correct, missing (false negative) or spurious (false positive)
- You can sort the columns however you like



# Comparing the annotations

Annotation Diff Tool

Key doc: ft-airlines-27-jul-200... Key set: Key Type: Organization Weight: 1.0

Resp. doc: ft-airlines-27-jul-200... Resp. set: [Default set] Features: all some none 1.0

Start	End	Key	Features	=?	Start	End	
1932	1936	Nats	{}	=	1932	1936	Nats
2456	2460	Nats	{}	=	2456	2460	Nats
2070	2075	LATCC	{}	=	2070	2075	LATCC
1354	1362	Barclays	{}	=	1354	1362	Barclays
1784	1788	Nats	{}	=	1784	1788	Nats
1751	1768	The·Airline·Group	{}	~	1755	1768	Airline·Gro
938	955	The·Airline·Group	{}	~	942	955	Airline·Gro
1669	1686	the·Airline·Group	{}	~	1673	1686	Airline·Gro
2412	2429	The·Airline·Group	{}	~	2416	2429	Airline·Gro
1266	1283	The·Airline·Group	{}	~	1270	1283	Airline·Gro
1052	1068	Monarch·Airlines	{}	~	1030	1068	Britannia·A
2029	2068	London·Area·and·Terminal·Control·Centre	{}	~	2045	2068	Terminal·C
634	640	Labour	{}	-?			
1030	1047	Britannia·Airways	{}	-?			
				?-	2029	2040	London·Are
				?-	2386	2395	Hampshire

Correct: 19    Recall Precision F-measure

Partially correct: 7    Strict: 0.68 0.68 0.68

Missing: 2    Lenient: 0.93 0.93 0.93

False positives: 2    Average: 0.80 0.80 0.80

10 documents loaded

Statistics    Adjudication

Key annotations    Response annotations



# Corpus Quality Assurance

---

- Corpus Quality Assurance tool extends the Annotation Diff functionality to the entire corpus, rather than on a single document at a time
- It produces statistics both for the corpus as a whole (Corpus statistics tab) and for each document separately (Document statistics tab)
- It compares two annotation sets, but makes no assumptions about which (if either) set is the gold standard. It just labels them A and B.
- This is because it can be used to measure Inter Annotator Agreement (IAA) where there is no concept of “correct” set



# Try out Corpus Quality Assurance

- Open your hands-on corpus and click the Corpus Quality Assurance tab at the bottom of the Display pane.

The screenshot shows the GATE Developer interface. The main window displays a list of documents in a corpus. The 'Corpus Quality Assurance' tab is selected at the bottom of the Display pane.

Index	Document name
0	ft-BT-07-aug-2001.xml_0001B
1	ft-BT-briefing-02-aug-2001.xml_0001C
2	ft-BT-loop-01-aug-2001.xml_0001D
3	ft-GKN-09-aug-2001.xml_0001E
4	ft-SSL-10-aug-2001.xml_0001F
5	ft-WestLB-BT-05-aug-2001.xml_00020
6	ft-airlines-27-jul-2001.xml_00021
7	ft-airtours-08-aug-2001.xml_00022
8	ft-bank-of-england-02-aug-2001.xml_00023
9	ft-bank-of-uk-08-Aug-2001.xml_00024
10	ft-bmi-09-may-2001.xml_00025
11	ft-bmi-25-feb-2001.xml_00026
12	ft-bmi-airline-07-aug-2001.xml_00027
13	ft-bt-03-aug-2001.xml_00028
14	ft-bt-26-jul-2001.xml_00029





# Select Annotation Sets

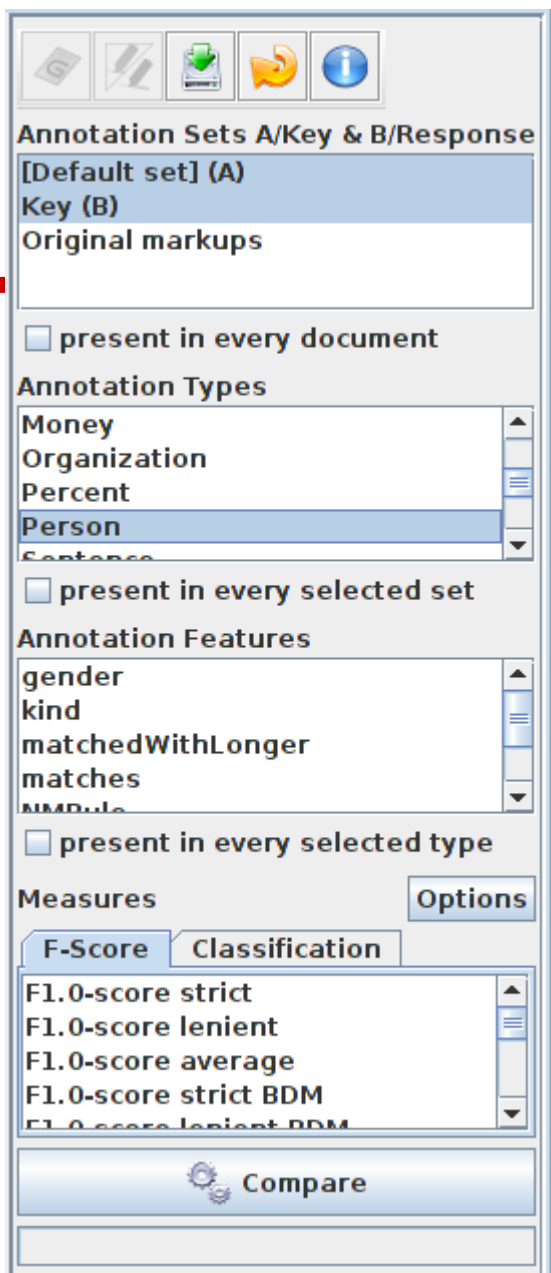
The screenshot shows the GATE software interface. The main window is titled 'Document statistics' and has tabs for 'Annotation', 'Match', 'Only A', 'Only B', and 'Overlap'. On the right side, a dialog box titled 'Annotation Sets A/Key &amp; B/Responses' is open. This dialog box has a red border and contains the following elements: a 'Default set' section with 'Key' and 'Original markups' listed; a checkbox 'present in every document'; an 'Annotation Types' section with a checkbox 'present in every selected set'; an 'Annotation Features' section with a checkbox 'present in every selected type'; a 'Measures' section with an 'Options' button and a list of measures including 'F1.0-score strict', 'F1.0-score lenient', 'F1.0-score average', 'F1.0-score strict BDM', and 'F1.0-score lenient BDM'; and a 'Compare' button at the bottom.

Select the annotation sets you wish to compare.

Click on the Key annotation set – this will label it set A.

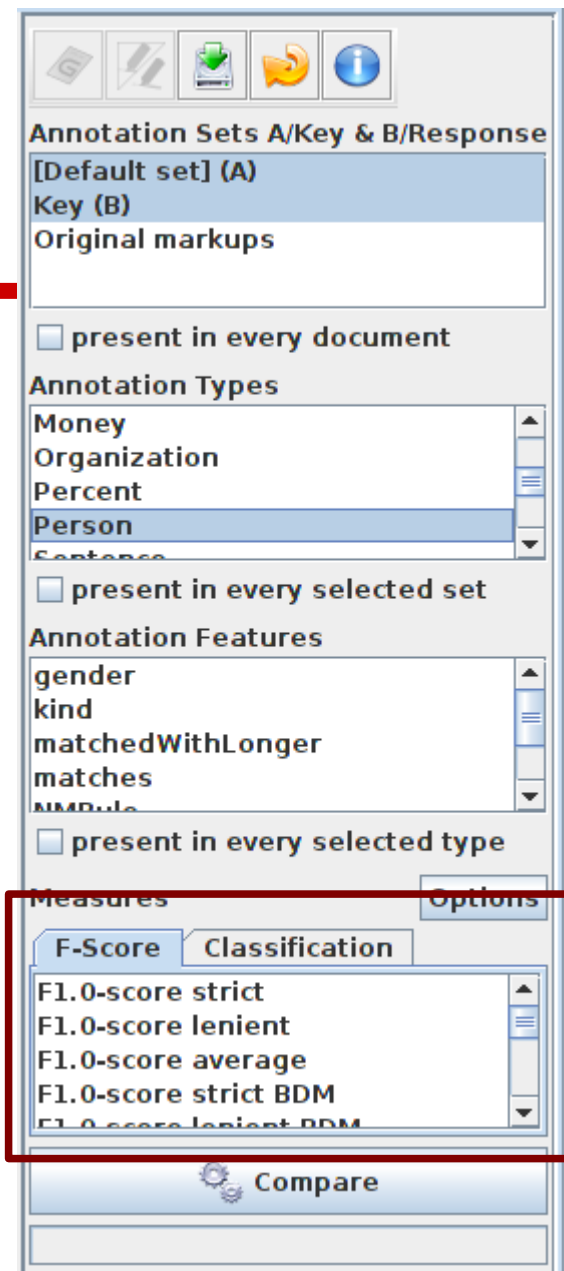
Now click on the default annotation set - this will label it set B.

# Select Type



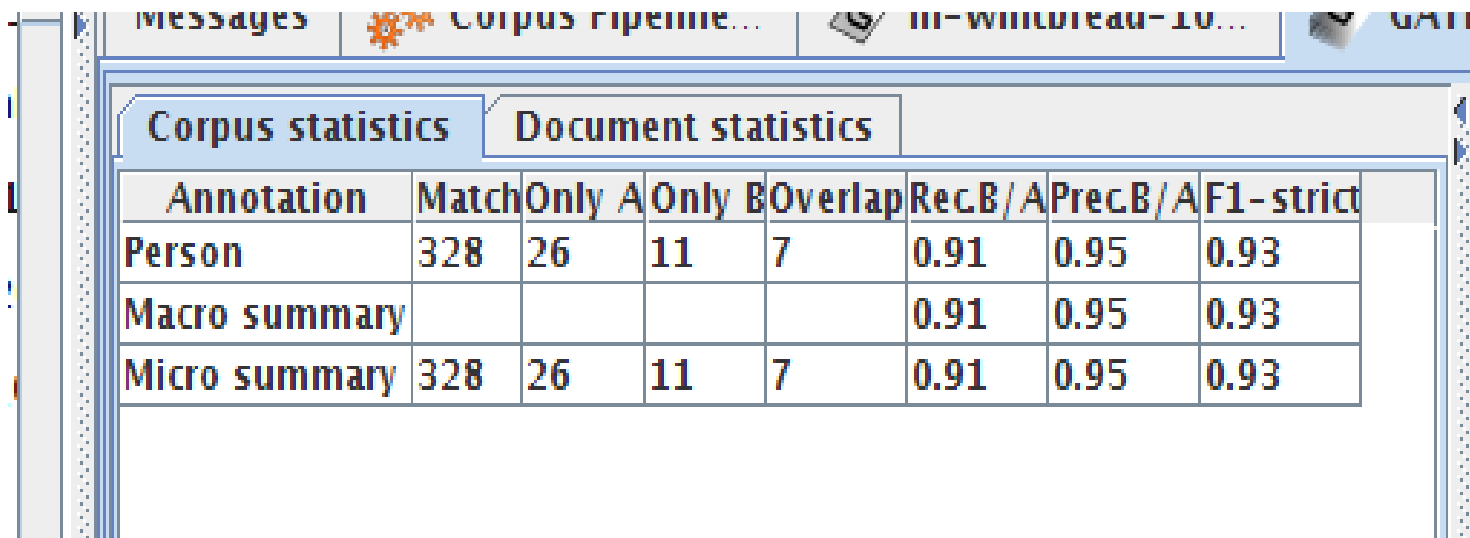
- Select the annotation type to compare (suggestion: select Organisation, Person and Location for now)
- Select the features to include (if any – leave unselected for now)
- You can select as many types and features as you want.

# Select measure



- In the “Measures” box, select the kind of F score you want “Strict, Lenient, Average” or any combination of them. Suggestion: try just “lenient” at first
- Select Compare

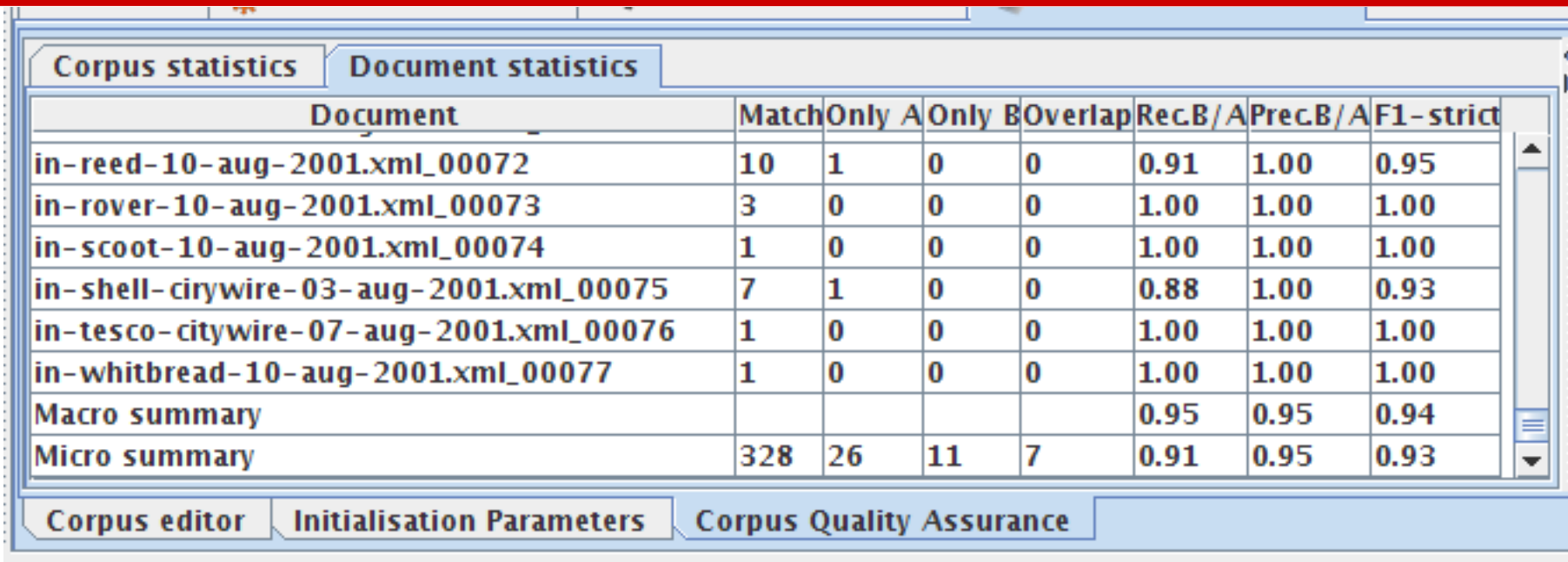
# Corpus Statistics Tab

A screenshot of the GATE software interface showing the 'Corpus statistics' tab. The window title bar includes 'messages', 'Corpus ripenne...', and 'in-winbread-10...'. The 'Corpus statistics' tab is active, displaying a table with columns for 'Annotation', 'Match', 'Only A', 'Only B', 'Overlap', 'Rec.B/A', 'Prec.B/A', and 'F1-strict'. The data rows are 'Person', 'Macro summary', and 'Micro summary'.

Annotation	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
Person	328	26	11	7	0.91	0.95	0.93
Macro summary					0.91	0.95	0.93
Micro summary	328	26	11	7	0.91	0.95	0.93

- Each annotation type is listed separately
- Precision, recall and F measure are given for each
- Two summary rows provide micro and macro averages

# Document Statistics Tab



Document	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
in-reed-10-aug-2001.xml_00072	10	1	0	0	0.91	1.00	0.95
in-rover-10-aug-2001.xml_00073	3	0	0	0	1.00	1.00	1.00
in-scoot-10-aug-2001.xml_00074	1	0	0	0	1.00	1.00	1.00
in-shell-citywire-03-aug-2001.xml_00075	7	1	0	0	0.88	1.00	0.93
in-tesco-citywire-07-aug-2001.xml_00076	1	0	0	0	1.00	1.00	1.00
in-whitbread-10-aug-2001.xml_00077	1	0	0	0	1.00	1.00	1.00
Macro summary					0.95	0.95	0.94
Micro summary	328	26	11	7	0.91	0.95	0.93

- Each document is listed separately
- Precision, recall and F measure are given for each
- Two summary rows provide micro and macro averages

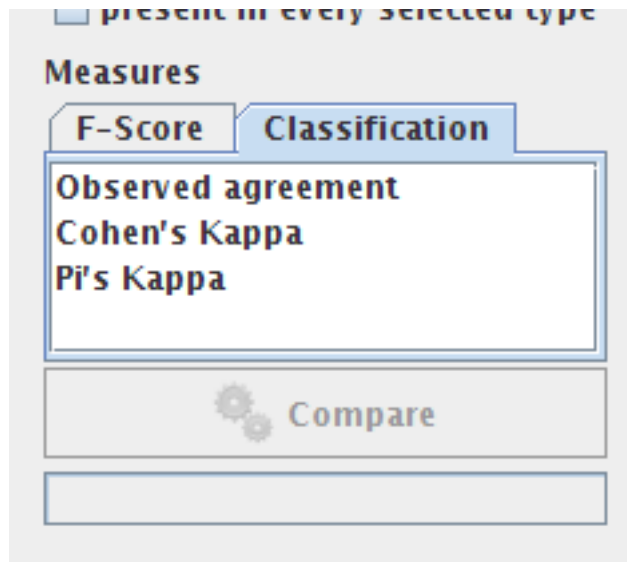


# Micro and Macro Averaging

---

- Micro averaging treats the entire corpus as one big document, for the purposes of calculating precision, recall and F
- Macro averaging takes the average of the rows

# Classification Measures



- By default, Corpus Quality Assurance presents the F-measures
- However, classification measures are also available
- These are not suitable for entity extraction tasks



# Corpus Quality Assurance PR

---

- Corpus QA can also be carried out as part of a GATE pipeline, using the Corpus QA PR
- The Corpus QA PR can be found in the tools plugin
- The PR writes out HTML pages, giving the same measures as the Corpus QA viewer
- The Corpus QA PR is executed when a pipeline reaches the last document in the corpus.
- You can set parameters for:
  - Annotation sets to use as key and response
  - Annotation types and features to compare
  - Evaluation metric to use





# Corpus Quality Assurance PR

---

- You must also set the URL of an output directory
- The PR writes HTML pages to this directory, giving the same measures as the Corpus QA viewer:
  - Per-document metrics
  - Corpus and annotation type metrics
- The output HTML is also linked to HTML generated by the Annotation Diff tool for each document
- You can thus use the PR to generate a full evaluation and click through to error reports for each document
- The extra exercises contains an example of running a pipeline with the Corpus QA PR

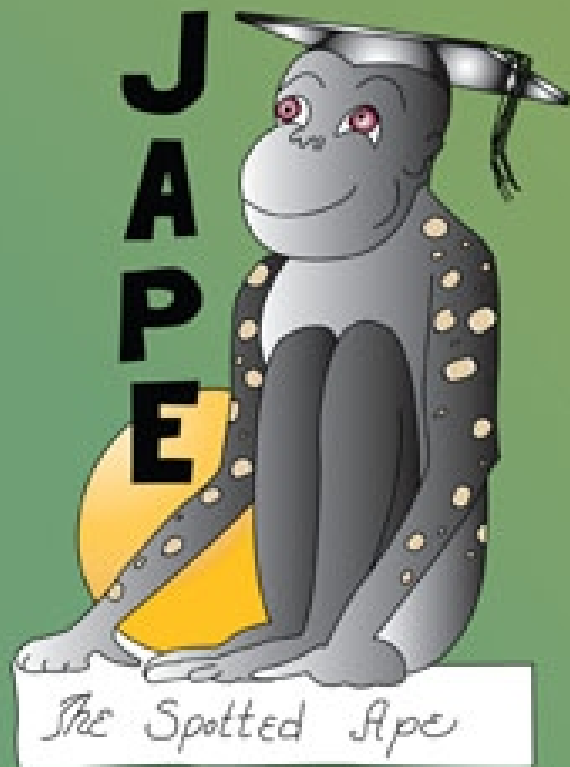


# Summary

---

- Module 2 has been devoted to IE and ANNIE
- You should now have a basic understanding of:
  - what IE is
  - how to load and run ANNIE
  - what each of the ANNIE components do
  - how to modify ANNIE components
  - multilingual capabilities of GATE
  - Evaluation

# Tomorrow: introducing **GATE** JAPE



By Mary Louise Janz  
Illustrated by Jared Ithaca

JAPE, a happy little ape, was always kind and thoughtful. His fine, bright mind helped him find his place in life with an unusual solution to his problem....



---

# Further exercises: sentence splitter variants



# Sentence Splitter variants

---

- Organisations do not span sentence boundaries, according to the rules used to create them.
- Load the default ANNIE and run it on the document in the directory `module2-hands-on/universities`
- Look at the Organisation annotations
- Now remove the sentence splitter and replace it with the alternate sentence splitter (see slide on Sentence Splitting variants for details)
- Run ANNIE again and look at the Organisation annotations.
- Can you see the difference?
- Can you understand why? If not, have a look at the relevant Sentence annotations.



---

# Further exercises: an ontology gazetteer



# Ontology Gazetteer

---

- This exercise opens a pre-configured application that contains an ontology based gazetteer, so that you can run it and look at the kind of results produced
- The exercise is not intended to explain ontologies or any of the ontology technologies used, and does not attempt to configure anything. These are covered in the advanced GATE course
- It is intended only to give you a flavour of what is possible
- GATE contains various ontology tools and gazetteers. We will use the Large Knowledge Base Gazetteer
- This is found in the Gazetteer\_LKB plugin



# Ontology Gazetteer

- You need a working internet connection for this exercise
- Restart GATE, or close all documents and PRs to tidy up
- Using the “File > Restore Application from file” menu, navigate to this directory in your GATE installation:
  - plugins\Gazeteer\_LKB\samples\sample\_pipelines
- Select and open this application file
  - sample\_linked\_data\_mashup.gapp





# Ontology Gazetteer

- The example application file contains a corpus pipeline with three PRs, and a corpus containing a single document from which to load it.
- Open the pipeline and take a look at the order of the PRs
  - The first PR is a Document Reset PR
  - The second is an LKB Gazetteer
- Double click on the LKB Gazetteer in the Processing Resources tree, to see its initialization parameters



# Ontology Gazetteer

---

- The LKB parameter `dictionaryPath` points to a directory that contains configuration files.
- These tell it where to find an ontology and how to use it. In our case, one of these points to an ontology at <http://factforge.net/sparql> and another contains a query to retrieve the names of actors from this ontology.
- When initialized, the PR builds a gazetteer from the results of the query. It can be configured to cache this gazetteer locally.
- When run, it will create Lookup annotations from this gazetteer, with features for classes and instances in the ontology.



# Ontology Gazetteer

- The third PR is a Semantic Enrichment PR
- **Look at its initialization parameters**
- The parameter repositoryUrl points to an ontology, in this case the same one as before - FactForge
- **Look at its runtime parameters in the pipeline view**
  - The parameter annotationTypes contains the single type Lookup
  - The parameter called query contains a query against the ontology
- The query will take ontology identifiers from Lookup annotations, look for their birthplace in FactForge, and add it to the annotation



# Ontology Gazetteer

---

- Run the pipeline over the corpus, and examine the annotations in the single document
- You should see Lookup annotations marking actors. Features are:
  - class, the URI of the class of Actor
  - inst, the URI of this particular actor
  - connections, URI of the actor's birthplace



---

# Further exercises: Quality Assurance PR



# Quality Assurance PR

---

- Corpus QA can also be carried out as part of a GATE pipeline, using the Quality Assurance PR.
- The PR writes out HTML pages, giving the same measures as the Corpus QA viewer.
- This exercise repeats the corpus evaluations from earlier in the tutorial, this time using the Quality Assurance PR
- The Quality Assurance PR can be found in the tools plugin
- **Restart GATE, or close all documents and PRs to tidy up**
- **Load the tools plugin, via the Plugin Management Console**



# Quality Assurance PR: preparation

---

- Create a new corpus and load it with the tutorial documents
- Take a look at the annotations.
- There is a set called “Key”. This is a set of annotations against which we want to evaluate ANNIE. In practice, they could be manual annotations, or annotations from another application.
- Load the ANNIE system with defaults, and open in the viewer
- **Important:** Change the runtime parameters for the Document Reset PR, adding “Key” to the setsToKeep parameter. This stops the application deleting our Key annotations when we run it.
- Create a new Quality Assurance PR
- Create an empty directory somewhere on your computer, into which results will be saved.



# Quality Assurance PR

---

- Add the Quality Assurance PR to the end of the pipeline
- Set parameters for:
  - keyASName set to Key
  - responseASName left blank to use the default set
  - Add the following to the annotationTypes list:
    - Organization
    - Person
    - Location
  - Evaluation metric to use, the “measure” parameter. Choose your preferred measure, e.g. F1\_STRICT





# Quality Assurance PR

---

- Set the QA PR's `outputFolderUrl` to the output directory that you created earlier
- Run the pipeline
- Examine the results in the output directory
  - `corpus-stats.html` shows the corpus statistics
  - `document-stats.html` shows the document statistics, and links to an annotation diff for each document and annotation type



---

**Further exercises:  
comparing ANNIE,  
LingPipe and OpenNLP**

# Comparing ANNIE, LingPipe and OpenNLP



- The idea of this exercise is to run and compare three different IE systems using the Corpus QA tools.
- As well as ANNIE, GATE includes wrappers for the independently developed NLP pipelines, LingPipe and OpenNLP
- All three systems are provided as pre-built applications through the GATE File menu
- Note that this is not a proper evaluation!
  - we are not using a gold standard
  - the three applications may have been built with different sets of guidelines

# Comparing ANNIE, LingPipe and OpenNLP



- Close any applications, documents and PRs that you have open in GATE
- Create a new corpus and populate it from the corpus in your tutorial material
- From the File → Ready Made Applications menu, load three applications:
  - ANNIE with defaults
  - LingPipe
  - OpenNLP

# Comparing ANNIE, LingPipe and OpenNLP



- We will compare the way in which the three applications create Person, Organization and Location annotations
- For comparison, we will need to put annotations from each application into a different annotation set. We will also need to normalize their names, so that each application creates annotations with exactly the same names
- We will do all of the above by using an Annotation Set Transfer PR at the end of each application. This is in the Tools plugin
- **Load the Tools plugin via the Plugin Management Console**



# ANNIE pipeline

- Create a new Annotation Set Transfer PR, calling it “annie transfer”
- Open the ANNIE application in the viewer
- Add “annie transfer” to the end and set parameters:
  - Set outputASName to “annie”
  - Add the following to the annotationTypes list, to copy these annotations:
    - Person
    - Organization
    - Location
- Select the first PR, the Document Reset PR, and add the following to the setsToKeep parameter list:
  - opennlp
  - lingpipe



# LingPipe pipeline

- Create a new Annotation Set Transfer PR, calling it “lingpipe transfer”
- Open the LingPipe application in the viewer
- Add “lingpipe transfer” to the end and set parameters:
  - Set outputASName to “lingpipe”
  - Add the following to the annotationTypes list , to copy and rename these annotations:
    - PERSON=Person
    - ORGANIZATION=Organization
    - LOCATION=Location
- Select the first PR, the Document Reset PR, and add the following to the setsToKeep parameter list:
  - opennlp
  - annie



# OpenNLP pipeline

- Create a new Annotation Set Transfer PR, calling it “opennlp transfer”
- Open the OpenNLP application in the viewer
- Add “opennlp transfer” to the end and set parameters:
  - Set outputASName to “opennlp”
  - Add the following to the annotationTypes list , to copy these annotations:
    - person=Person
    - organization=Organization
    - location=Location
- Select the first PR, the Document Reset PR, and add the following to the setsToKeep parameter list:
  - annie
  - lingpipe



# Comparing ANNIE, LingPipe and OpenNLP



- 
- Run each of the three applications over your corpus
  - Open the Corpus QA view, and do pair-wise comparisons of the three annotation sets, for the three annotation types
  - Look at the Document statistics tab, and open individual documents that differ
  - How do the three applications differ?

# More exercises with ANNIE, LingPipe and OpenNLP



- It is possible to mix the different PRs from the three applications, e.g. to replace the tokeniser of one with the tokeniser from another
- This doesn't always work – sometimes there are dependencies not met by equivalent PRs in the other applications
- The GATE documentation for the OpenNLP and LingPipe plugins has some notes on this
- For further exercises, you could try comparing the annotations output by individual PRs from each application
- You could also see what effect mixing PRs from different applications has on the final entity annotations