Module 2: Introduction to IE and ANNIE

About this tutorial

This tutorial comprises the following topics:

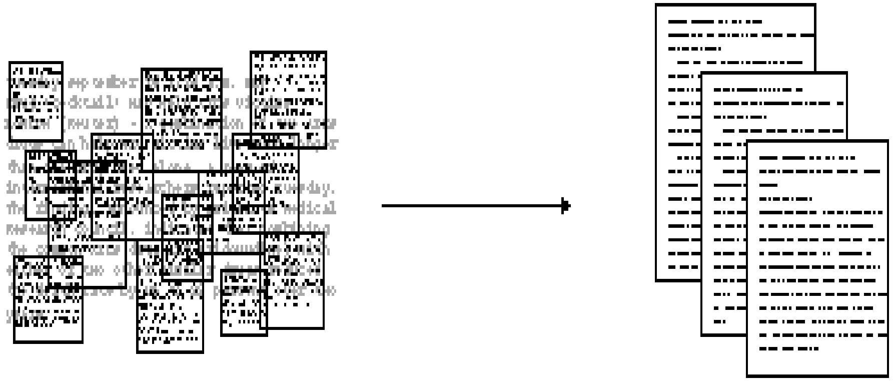
- Introduction to IE
- ANNIE
- Multilingual tools in GATE
- Evaluation and Corpus Quality Assurance

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

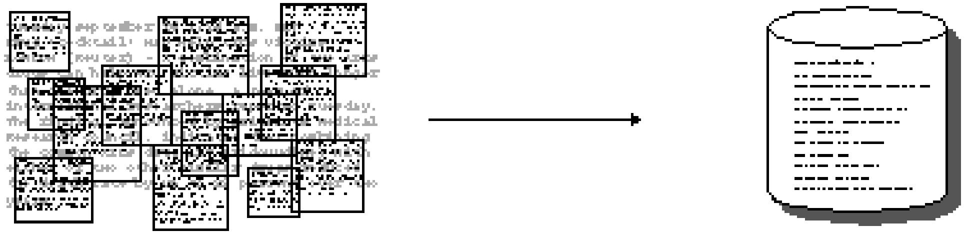
What is information extraction?

IE is not IR

- 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 thess?
- 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 as much 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, NER 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 etc.

Why is NER important?

- NER provides a foundation for building 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”

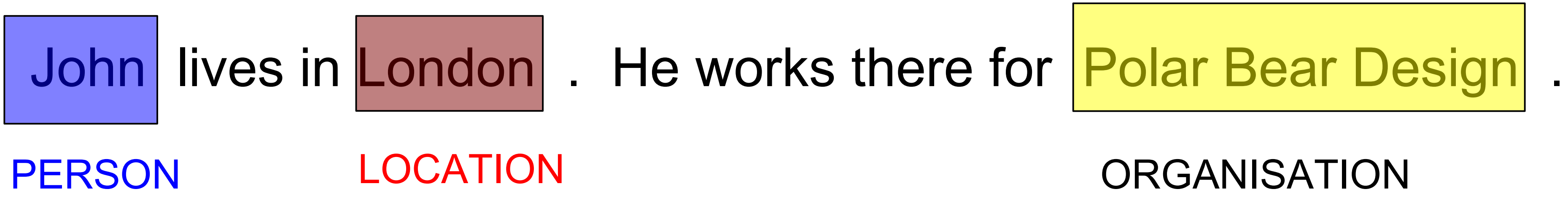
Typical NE pipeline

- 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



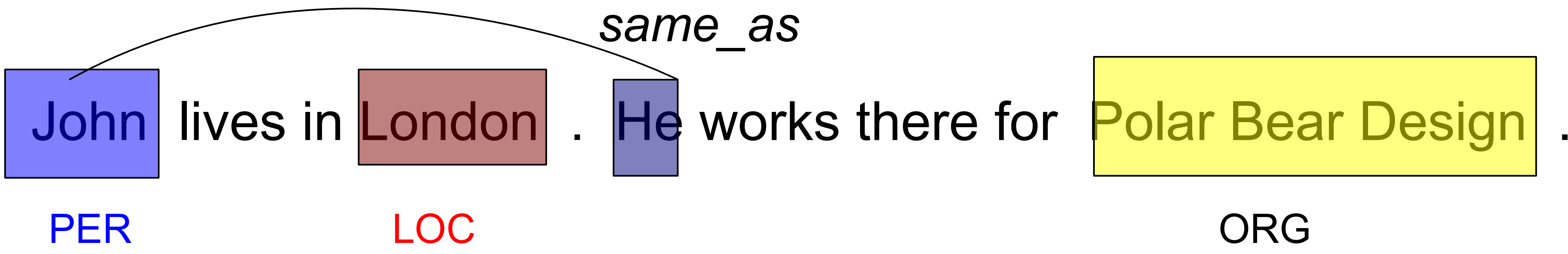
John lives in London . He works there for Polar Bear Design .

PERSON

LOCATION

ORGANISATION

Co-reference



same_as

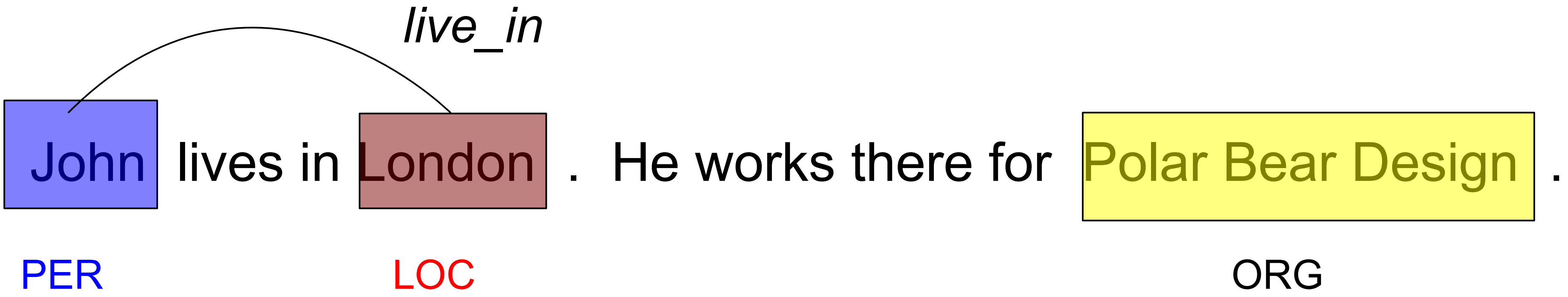
John lives in London . He works there for Polar Bear Design .

PER

LOC

ORG

Relations

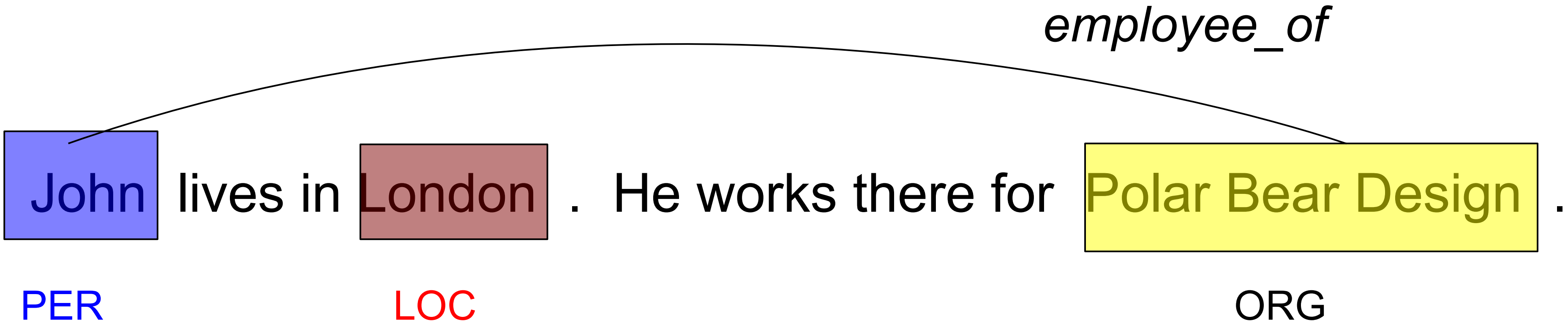


live_in

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

PER LOC ORG

Relations (2)

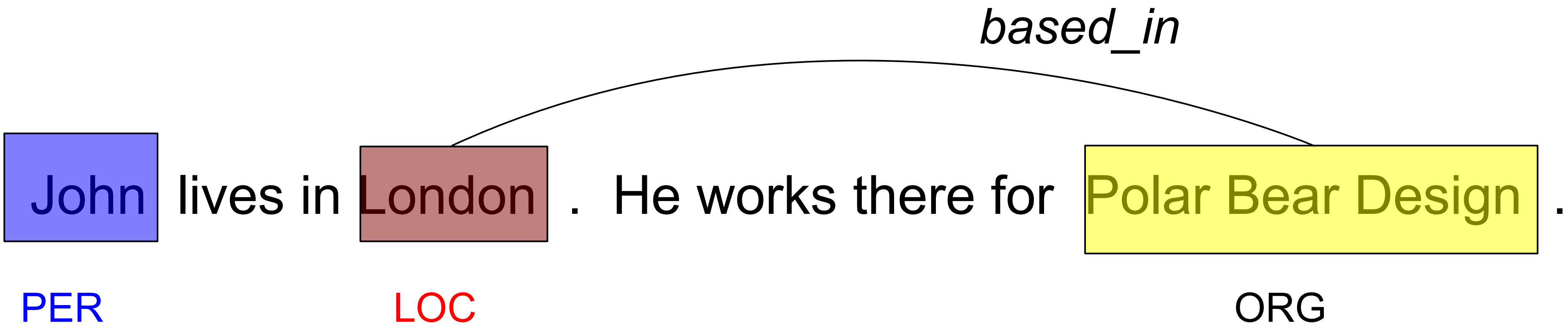


employee_of

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

PER LOC ORG

Relations (3)



based_in

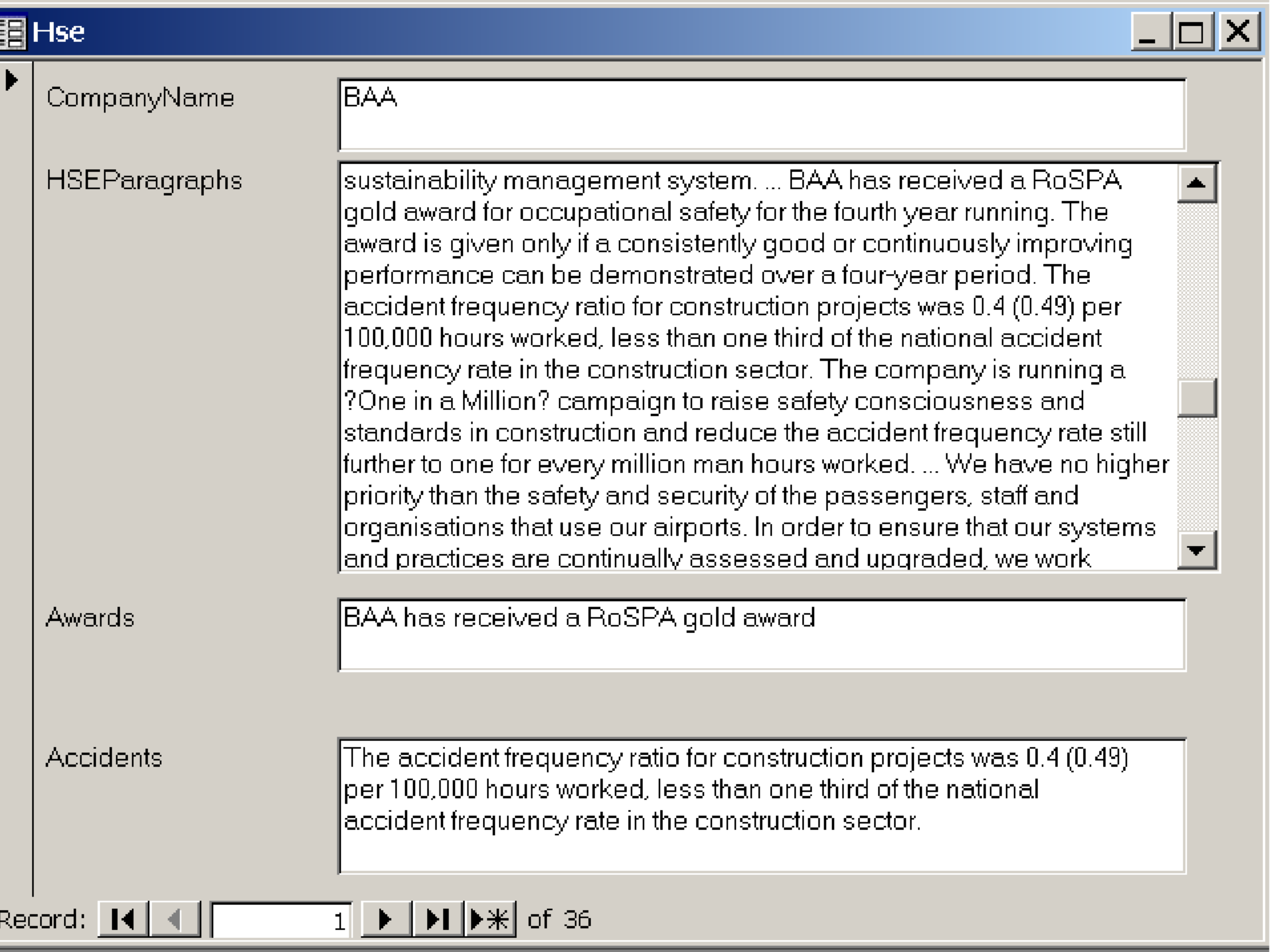
John lives in London . He works there for Polar Bear Design .

PER LOC ORG

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

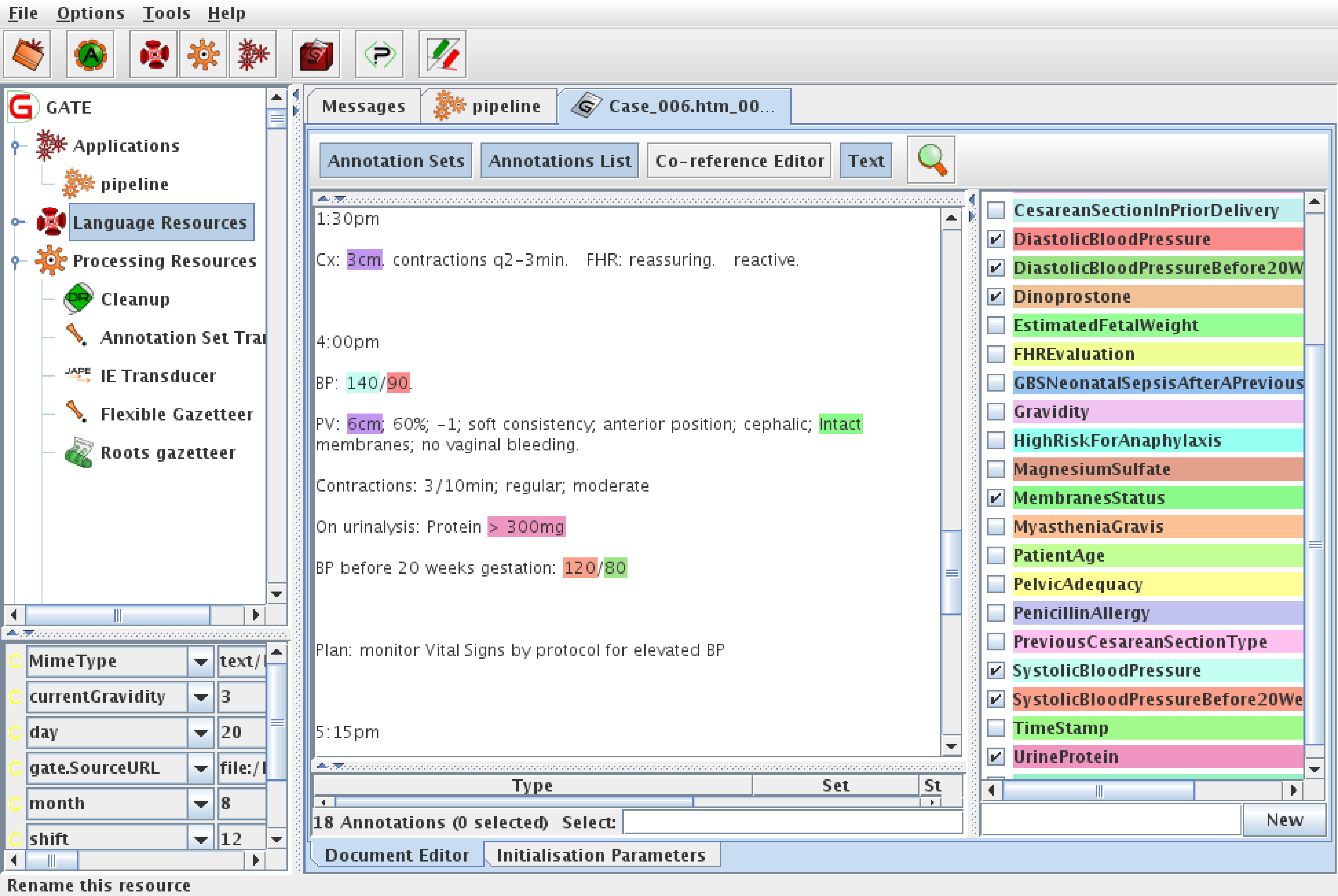
HSEParagraphs: sustainability management system. ... BAA has received a RoSPA gold award for occupational safety for the fourth year running. The award is given only if a consistently good or continuously improving performance can be demonstrated over a four-year period. The accident frequency ratio 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



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

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

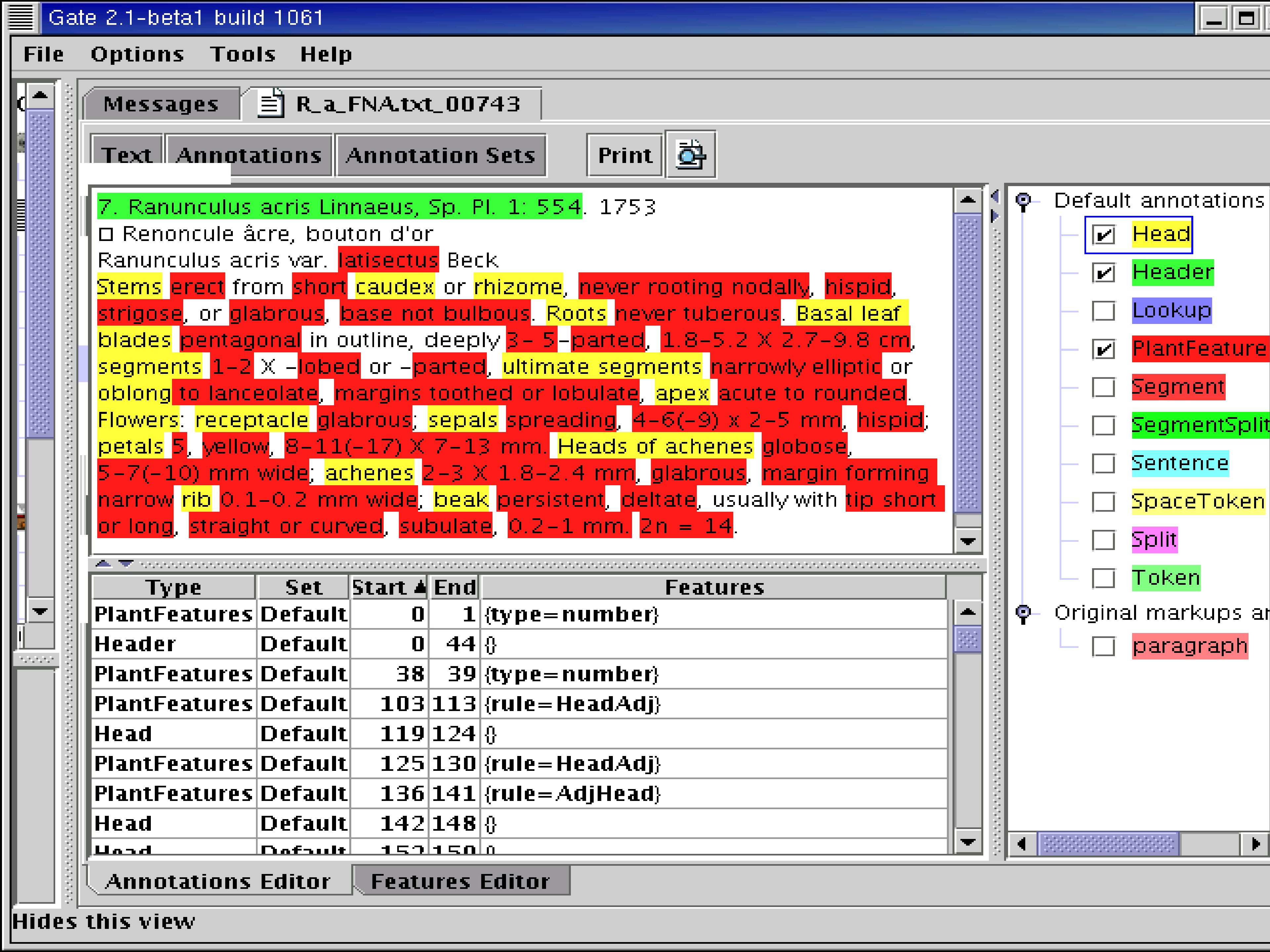
- □ CesareanSectionInPriorDelivery
- ☑ DiastolicBloodPressure
- ☑ DiastolicBloodPressureBefore20W
- ☑ Dinoprostone
- □ EstimatedFetalWeight
- □ FHREvaluation
- □ GBSNeonatalSepsisAfterAPrevious
- □ Gravidity
- □ HighRiskForAnaphylaxis
- □ MagnesiumSulfate
- ☑ MembranesStatus
- □ MyastheniaGravis
- □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

18 Annotations (0 selected)

Rename this resource

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.



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.

Default annotations
- ☑ Head
- ☑ Header
- □ Lookup
- ☑ PlantFeature
- □ Segment
- □ SegmentSplit
- □ Sentence
- □ SpaceToken
- □ Split
- □ Token

Original markups an
- □ paragraph

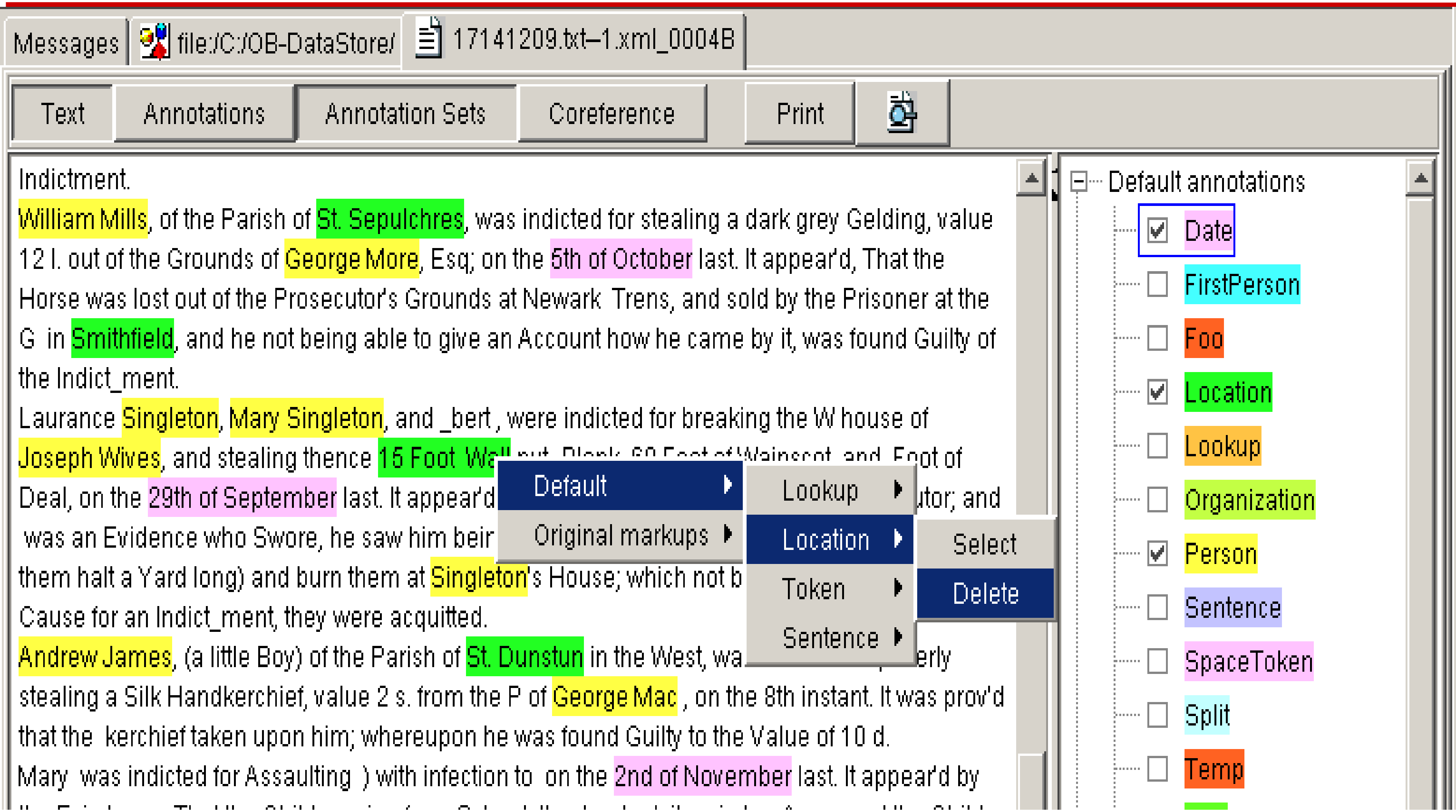
Type	Set	Start	End	Features
PlantFeatures	Default	0	1	{type=number}
Header	Default	0	44	{}
PlantFeatures	Default	38	39	{type=number}
PlantFeatures	Default	103	113	{rule=HeadAdj}
Head	Default	119	124	{}
PlantFeatures	Default	125	130	{rule=HeadAdj}
PlantFeatures	Default	136	141	{rule=AdjHead}
Head	Default	142	148	{}
Head	Default	152	158	{}

Hides this view

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

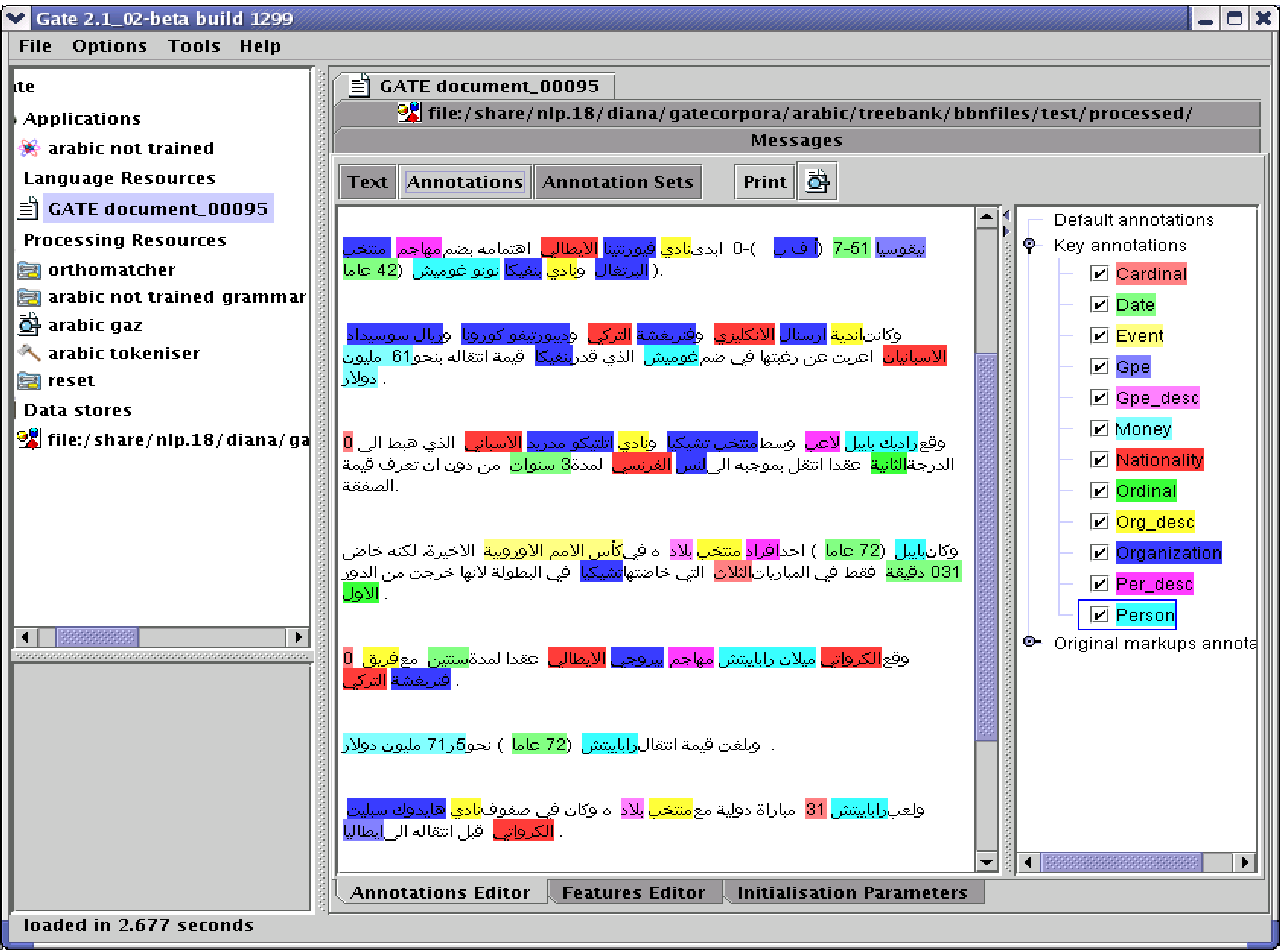


Indictment.
William Mills, of the Parish of St. Sepulchres, was indicted for stealing a dark grey Gelding, value
12 l. out of the Grounds of George More, Esq; on the 5th of October last. It appear'd, That the
Horse was lost out of the Prosecutor's Grounds at Newark Trens, and sold by the Prisoner at the
G in Smithfield, and he not being able to give an Account how he came by it, was found Guilty of
the Indict_ment.
Laurance Singleton, Mary Singleton, and _bert , were indicted for breaking the W house of
Joseph Wives, and stealing thence 15 Foot Wall out Plank, 60 Foot of Wainscot, and Foot of
Deal, on the 29th of September last. It appear'd utor; and
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 Indict_ment, they were acquitted.
Andrew James, (a little Boy) of the Parish of St. Dunstun in the West, wa erly
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 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



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

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

وقع راديك بايبل لاعب وسط منتخب تشيكيا ونادي اتلتيكو مدريد الاسباني الذي هبط الى 0 الدرجة الثانية عقدا انتقل بموجبه الى لنس الفرنسي لمدة 3 سنوات من دون ان تعرف قيمة الصفقة.

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

وقع الكرواتي ميلان رابايتش مهاجم بيروجيا الايطالي عقدا لمدة سنتين مع فريق 0 فنربغشة التركي .

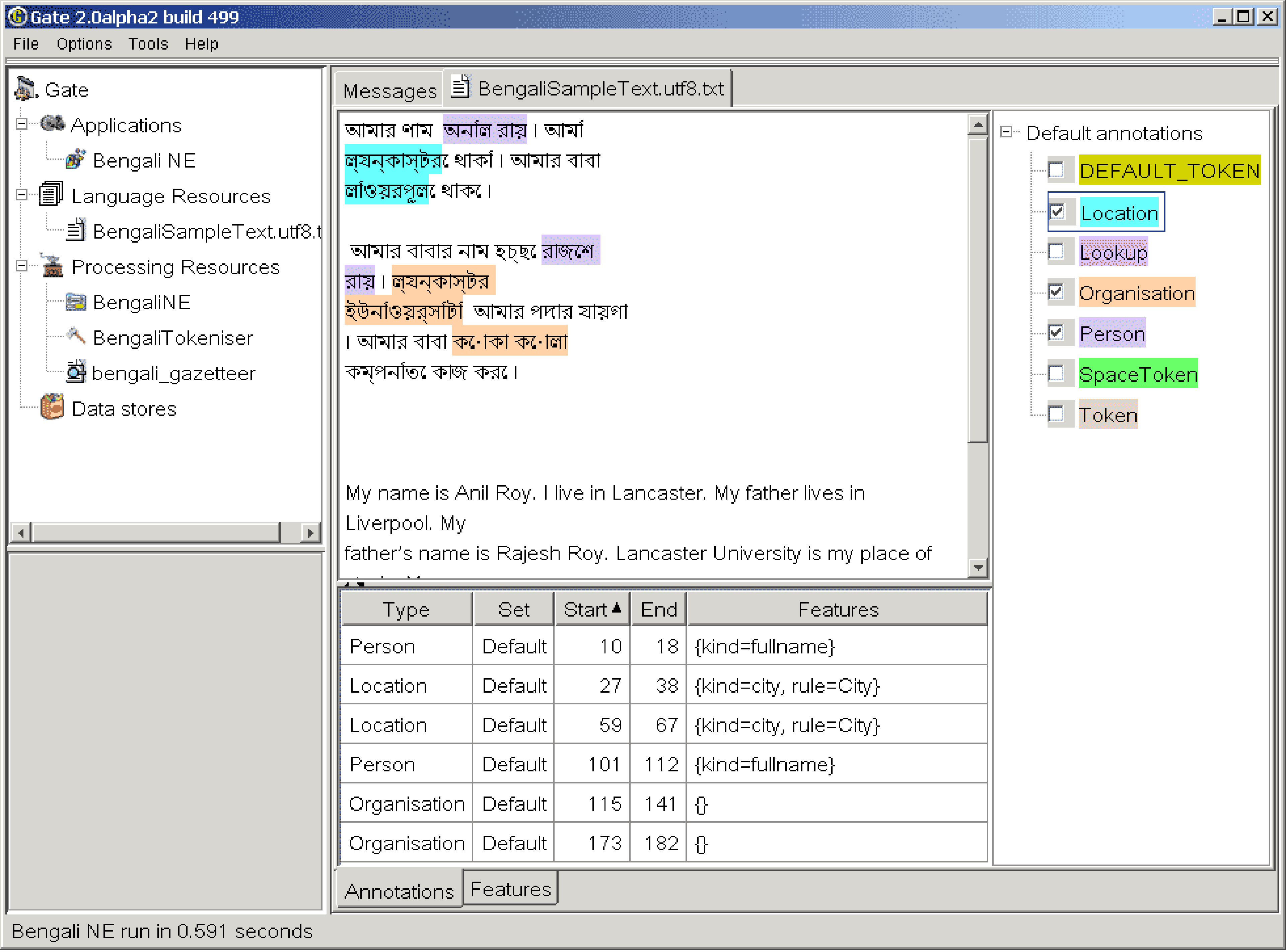
. وبلغت قيمة انتقال رابايتش (72 عاما) نحو 5ر71 مليون دولار

ولعب رابايتش 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

loaded in 2.677 seconds



আমার ণাম অনিল রায়। আমি
ল্যন্কাস্টরে থাকি। আমার বাবা
লিভয়রপুলে থাকে।

আমার বাবার নাম হচ্ছে রাজেশ
রায়। ল্যন্কাস্টর
ইউনিভয়র্সিটি আমার পদার যায়গা
। আমার বাবা কোকা কোলা
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	{}

Bengali NE run in 0.591 seconds

ANNIE: A Nearly New Information Extraction system

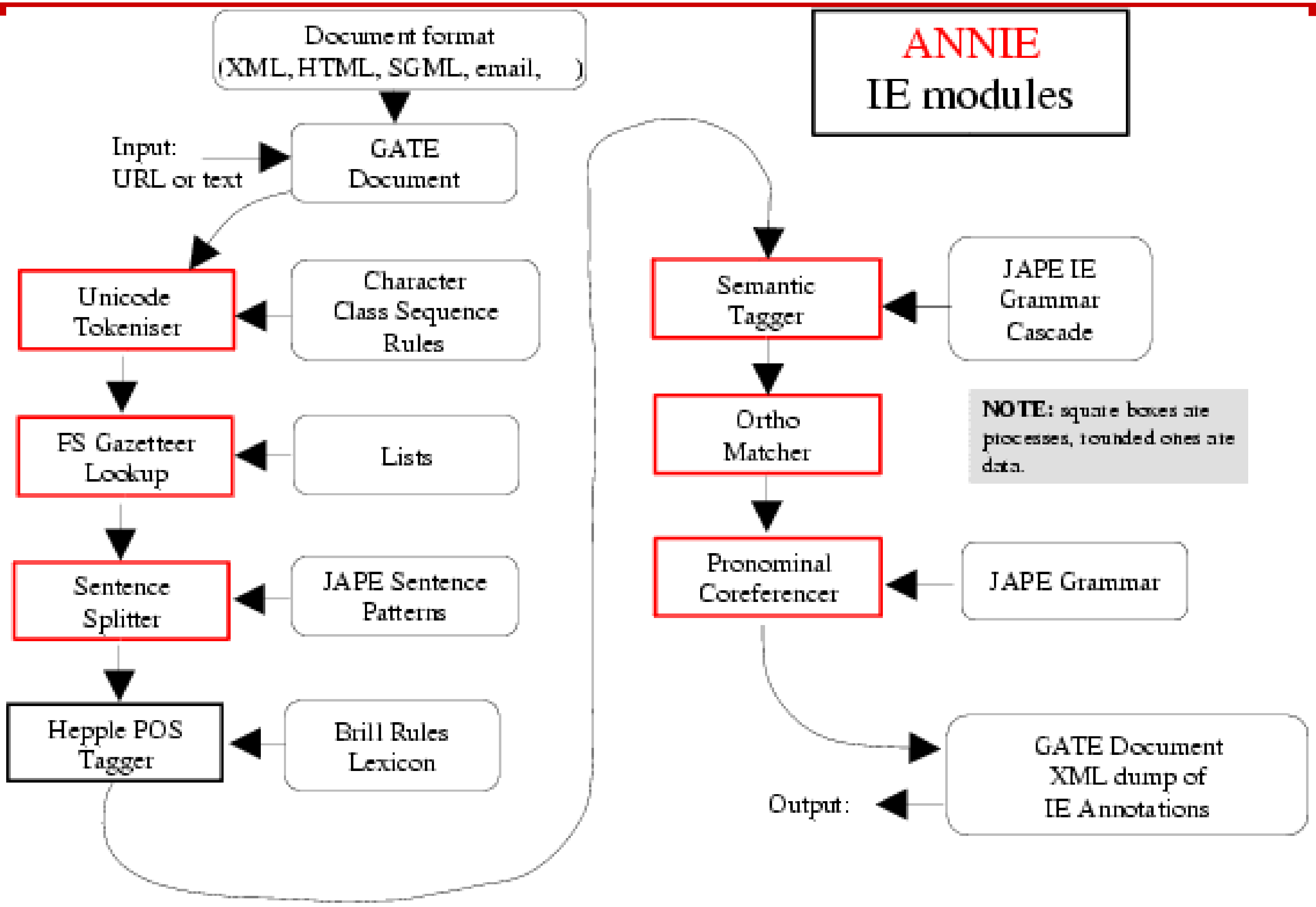
Nearly New Information Extraction

- ANNIE is a readymade 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



NOTE: square boxes are processes, rounded ones are data.

Loading and running ANNIE

- Because ANNIE is a ready-made application, we can just load it directly from the menu
- Click the A icon from the top GATE menu OR Select File → Load ANNIE system
- Select “with defaults”
- Load any document from module-2-hands-on/news-texts and add it to a corpus
- 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)

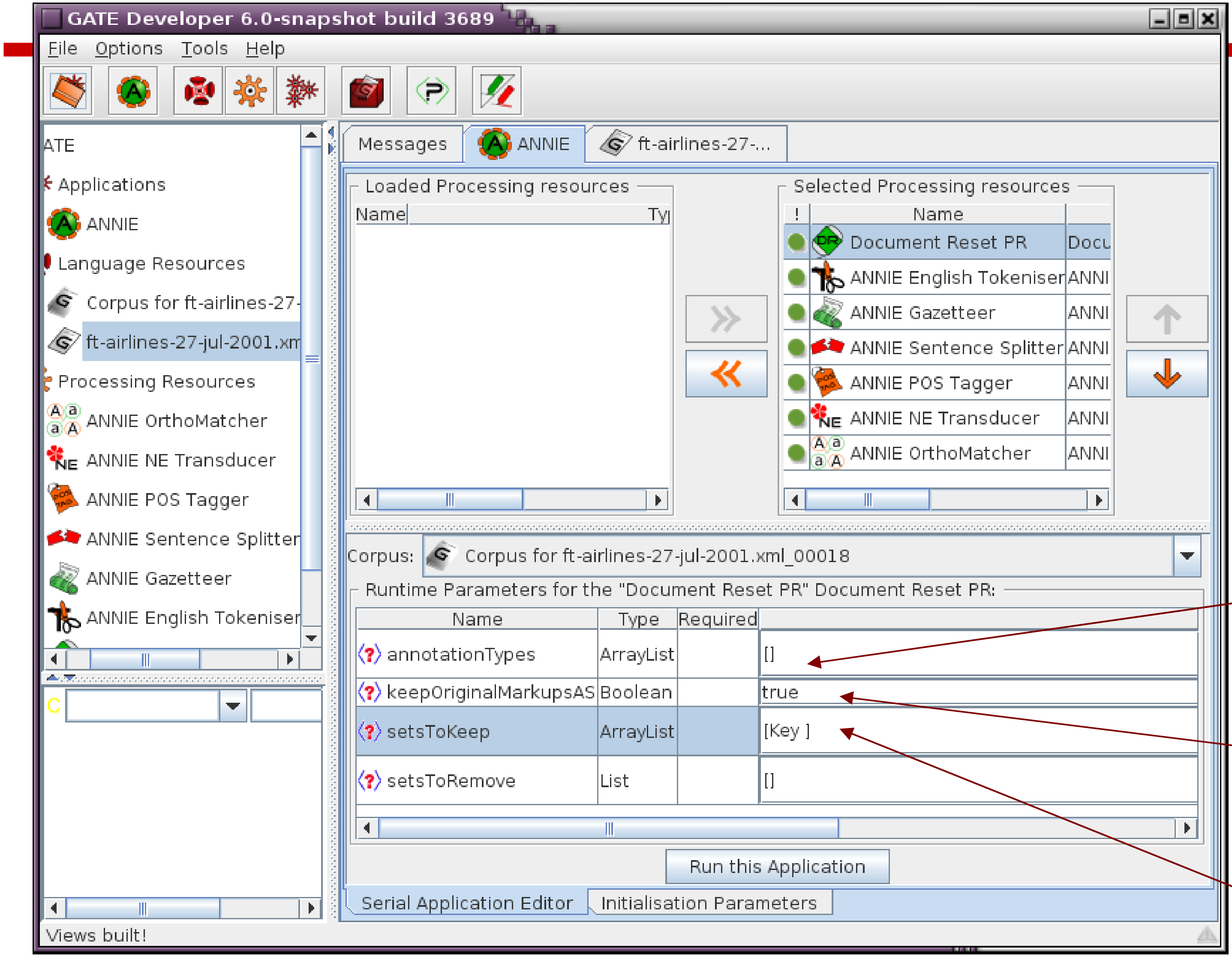
Let's look at the PRs

- Each PR in the ANNIE pipeline creates some new annotations, or modifies existing ones
- Document Reset → removes annotations
- Tokeniser → Token annotations
- Sentence Splitter → Sentence, Split annotations
- Gazetteer → Lookup annotations
- POS tagger → adds category features to Token annotations
- JAPE 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
- By default it does not remove the “Original markups” AS
- 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	Ty

Selected Processing resources

!	Name	
●	Document Reset PR	Docu
●	ANNIE English Tokeniser	ANNI
●	ANNIE Gazetteer	ANNI
●	ANNIE Sentence Splitter	ANNI
●	ANNIE POS Tagger	ANNI
●	ANNIE NE Transducer	ANNI
●	ANNIE OrthoMatcher	ANNI

Corpus: Corpus for ft-airlines-27-jul-2001.xml_00018

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

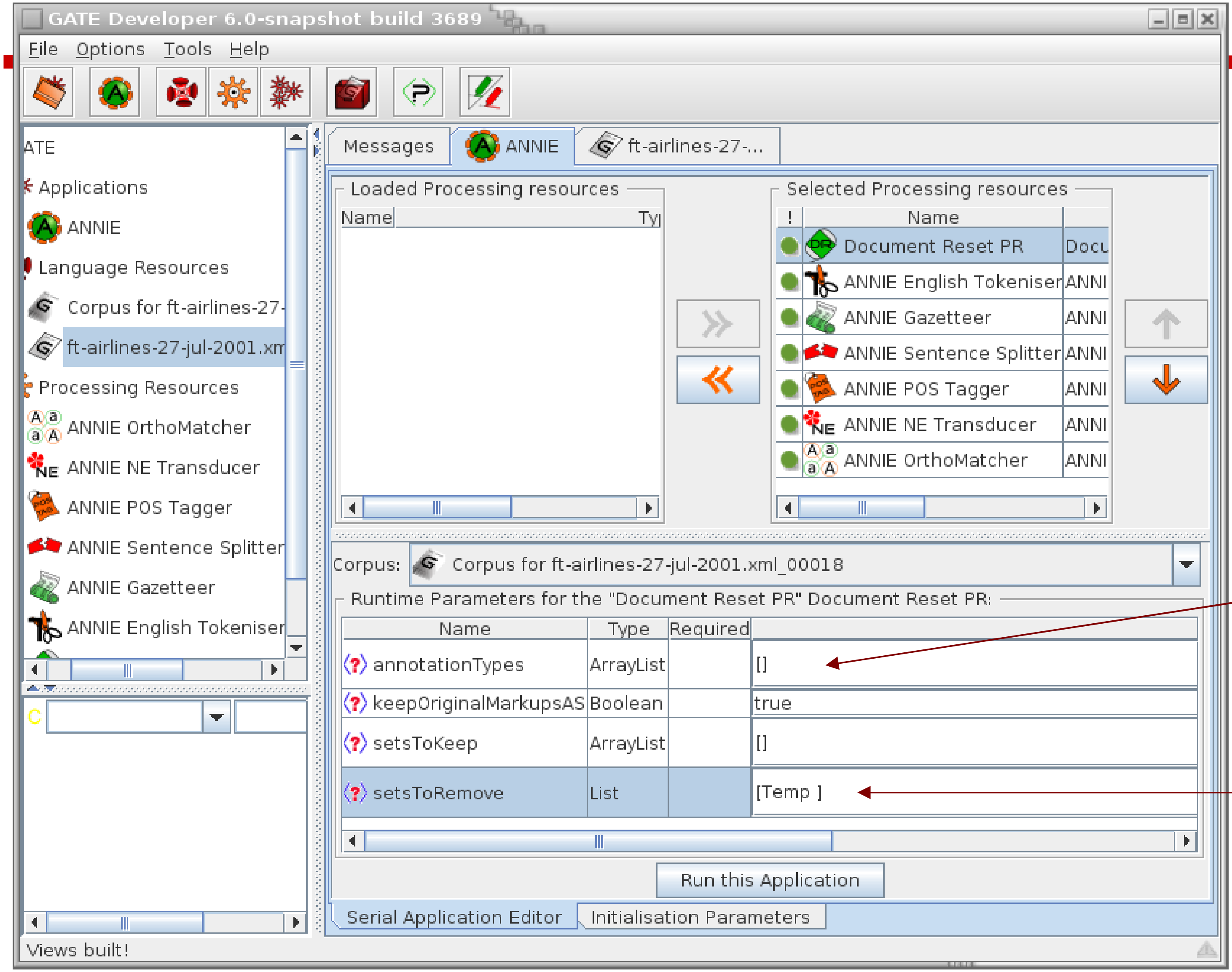
Name	Type	Required	
annotationTypes	ArrayList		[]
keepOriginalMarkupsAS	Boolean		true
setsToKeep	ArrayList		[Key]
setsToRemove	List		[]

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

Keep Original Markups set

Keep Key set

Document Reset Parameters



Selected Processing resources

!	Name	
●	Document Reset PR	Docu
●	ANNIE English Tokeniser	ANNI
●	ANNIE Gazetteer	ANNI
●	ANNIE Sentence Splitter	ANNI
●	ANNIE POS Tagger	ANNI
●	ANNIE NE Transducer	ANNI
●	ANNIE OrthoMatcher	ANNI

Corpus: Corpus for ft-airlines-27-jul-2001.xml_00018

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

Name	Type	Required	
annotationTypes	ArrayList		[]
keepOriginalMarkupsAS	Boolean		true
setsToKeep	ArrayList		[]
setsToRemove	List		[Temp]

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

Remove the Temp set only

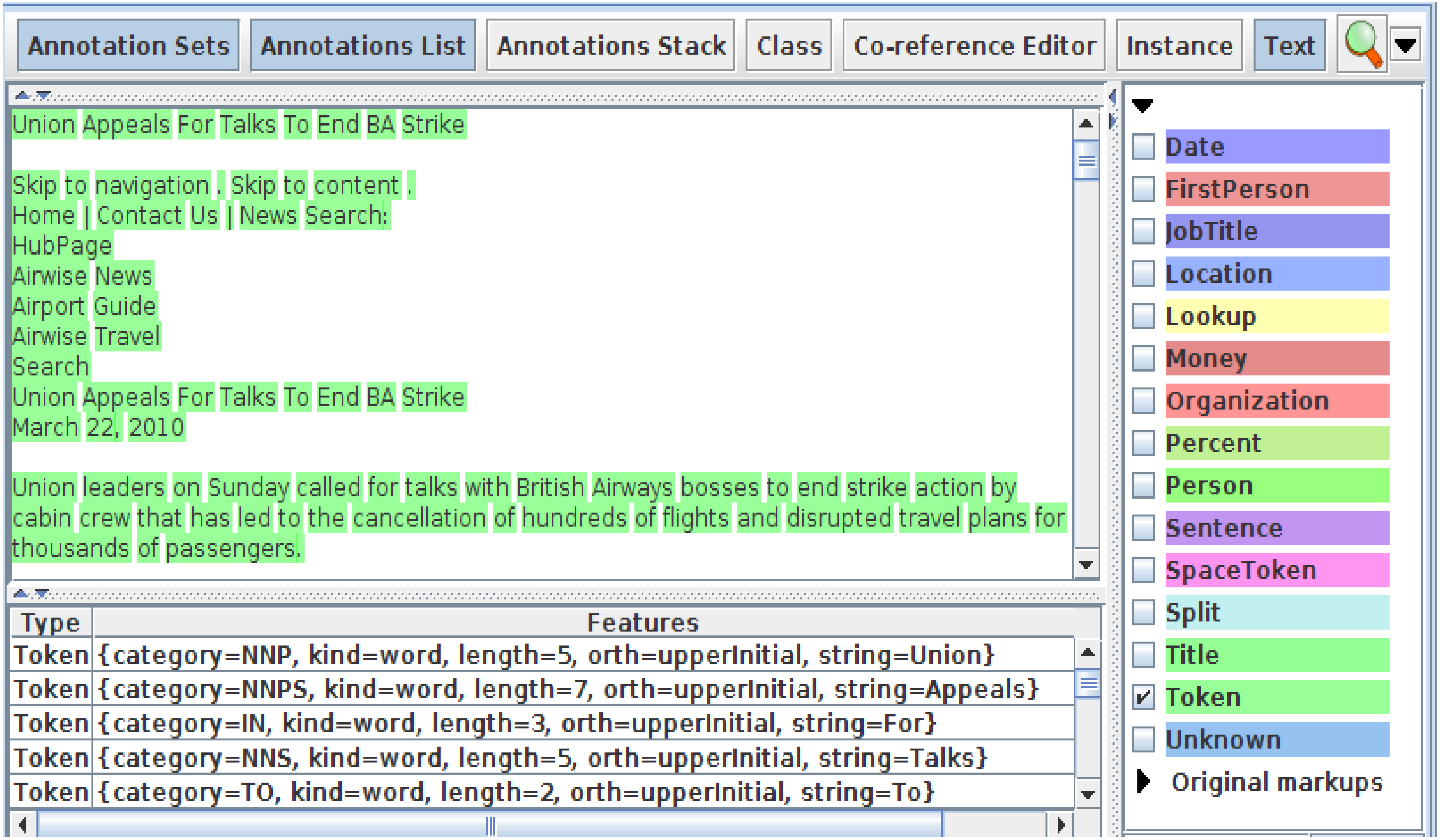
Tokenisation and sentence splitting

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



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

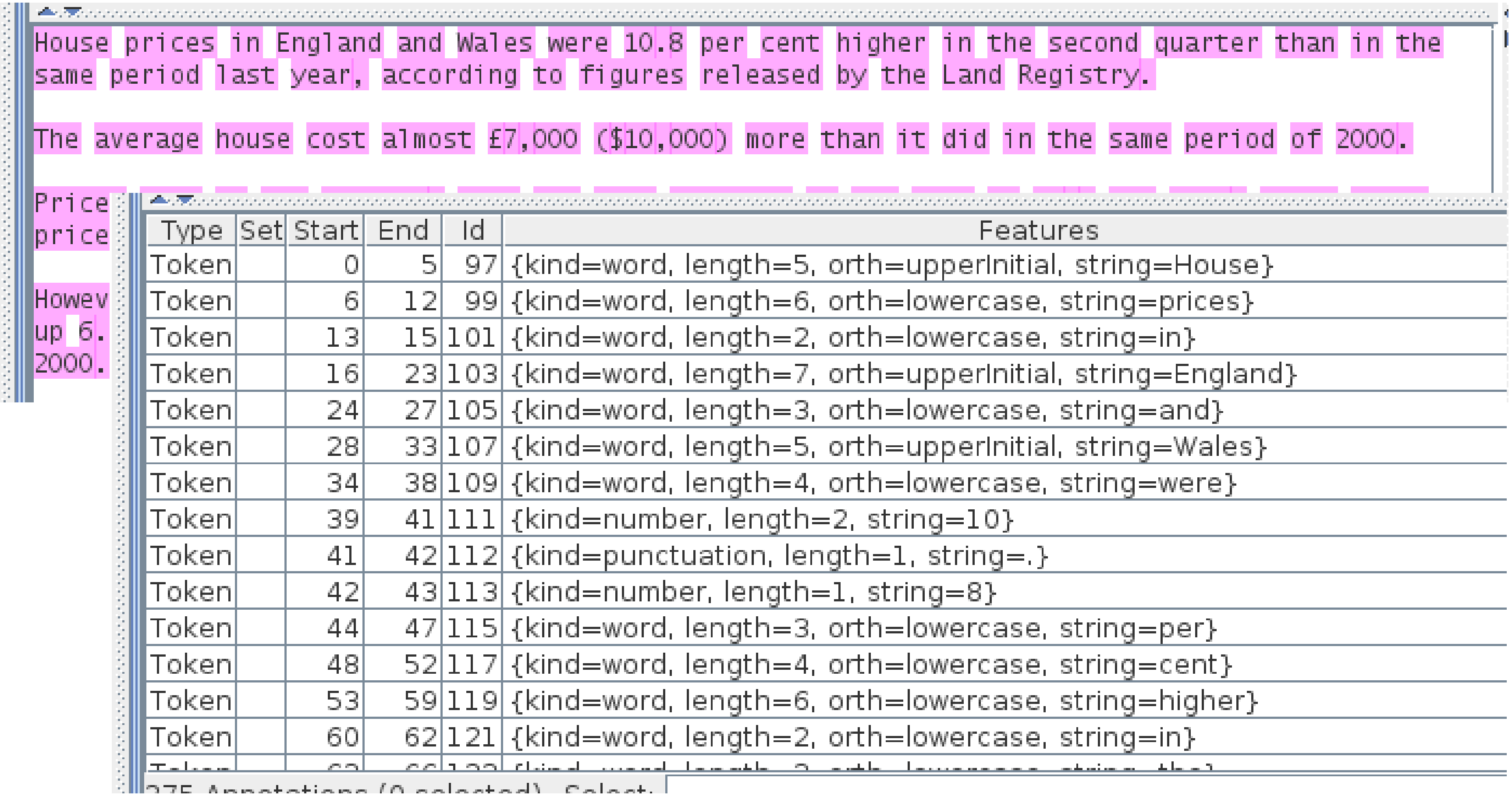
ANNIE English Tokeniser

- 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 (like the Penn Treebank)
  - 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)
- Create a new corpus and populate it with module
- Create a new application (corpus pipeline)
- Load a Document Reset and an ANNIE English Tokeniser
- Add them (in that order) to the application
- Set the Document Reset to keep the Key AS
- Run the application on the corpus
- View the Token and SpaceToken annotations
- What different values of the “kind” feature do you see?

Looking at Tokens



House prices in England and Wales were 10.8 per cent higher in the second quarter than in the same period last year, according to figures released by the Land Registry.

The average house cost almost £7,000 ($10,000) more than it did in the same period of 2000.

Price
price

Howev
up 6.
2000.

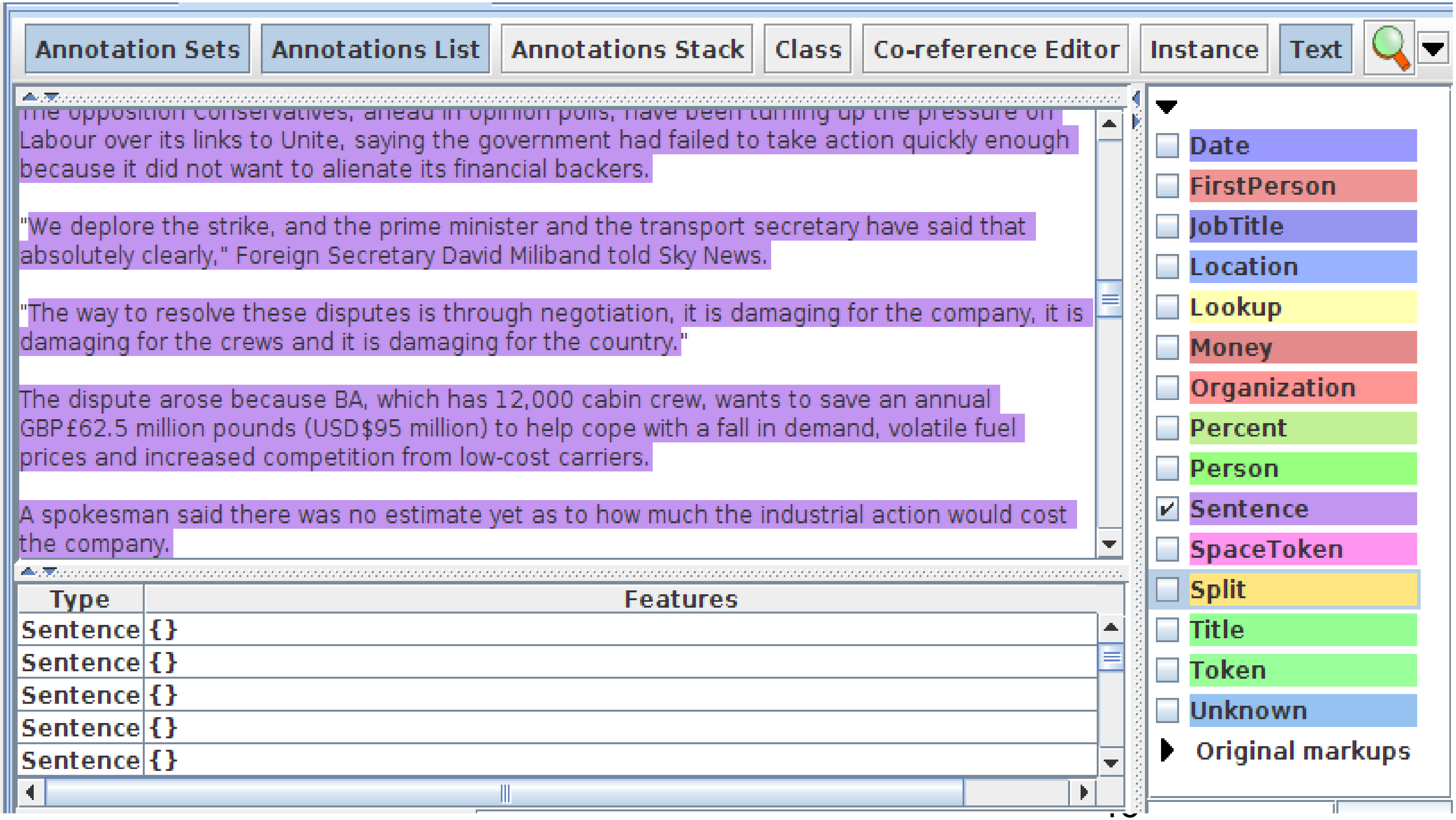
Type	Set	Start	End	Id	Features
Token		0	5	97	{kind=word, length=5, orth=upperInitial, string=House}
Token		6	12	99	{kind=word, length=6, orth=lowercase, string=prices}
Token		13	15	101	{kind=word, length=2, orth=lowercase, string=in}
Token		16	23	103	{kind=word, length=7, orth=upperInitial, string=England}
Token		24	27	105	{kind=word, length=3, orth=lowercase, string=and}
Token		28	33	107	{kind=word, length=5, orth=upperInitial, string=Wales}
Token		34	38	109	{kind=word, length=4, orth=lowercase, string=were}
Token		39	41	111	{kind=number, length=2, string=10}
Token		41	42	112	{kind=punctuation, length=1, string=.}
Token		42	43	113	{kind=number, length=1, string=8}
Token		44	47	115	{kind=word, length=3, orth=lowercase, string=per}
Token		48	52	117	{kind=word, length=4, orth=lowercase, string=cent}
Token		53	59	119	{kind=word, length=6, orth=lowercase, string=higher}
Token		60	62	121	{kind=word, length=2, orth=lowercase, string=in}
Token		63	66	123	{kind=word, length=3, orth=lowercase, string=the}

275 Annotations (0 selected) Select:

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 and add it to your application (after the tokeniser)
- Run the application and view the results

Document with Sentences



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

Sentence splitter variants

- An alternate set of rules can be loaded with the regular sentence splitter
- To do this, load “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 the Wall Street Journal corpus (news articles), uses Penn Treebank tagset
- Default ruleset and lexicon can be modified manually (with a little deciphering)
- Adds a category feature to Token annotations
- Requires Tokeniser and Sentence Splitter to be run first

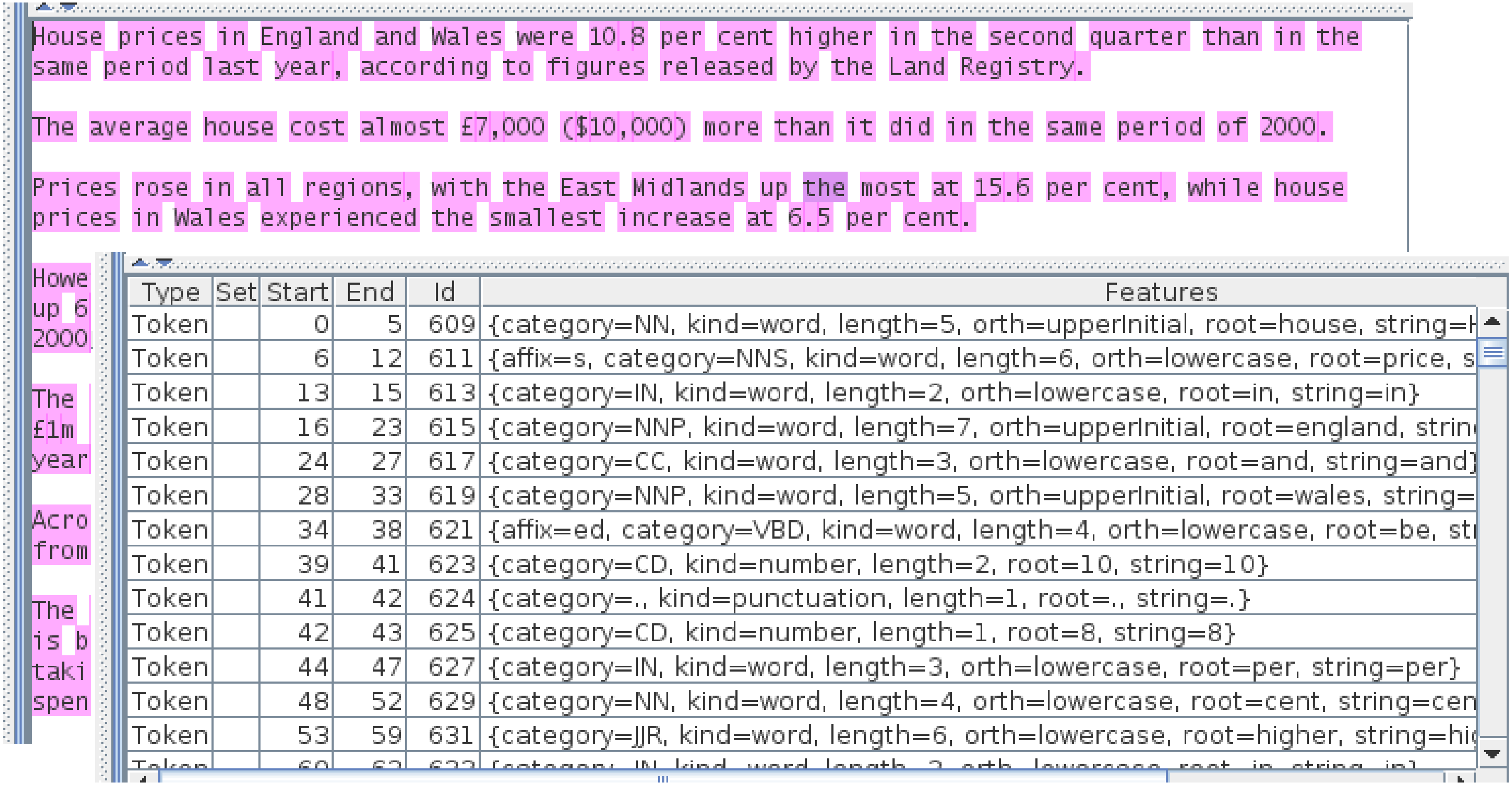
Morphological analyser

- This is in the Tools plugin (not an integral part of ANNIE)
- Flex based rules: can be modified by the user (instructions in the User Guide)
- Generates “root” feature on Token annotations
- The tokeniser must be run before the morpher
- The POS tagger must run before the morpher if the *considerPOSTag* parameter is set to true (and the POS tagger requires a sentence splitter first)

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
- Examine the features of the Token annotations
    - New features of category and root have been added

Shallow lexico-syntactic features



House prices in England and Wales were 10.8 per cent higher in the second quarter than in the same period last year, according to figures released by the Land Registry.

The average house cost almost £7,000 ($10,000) more than it did in the same period of 2000.

Prices rose in all regions, with the East Midlands up the most at 15.6 per cent, while house prices in Wales experienced the smallest increase at 6.5 per cent.

Howe
up 6
2000

The
£1m
year

Acro
from

The
is b
taki
spen

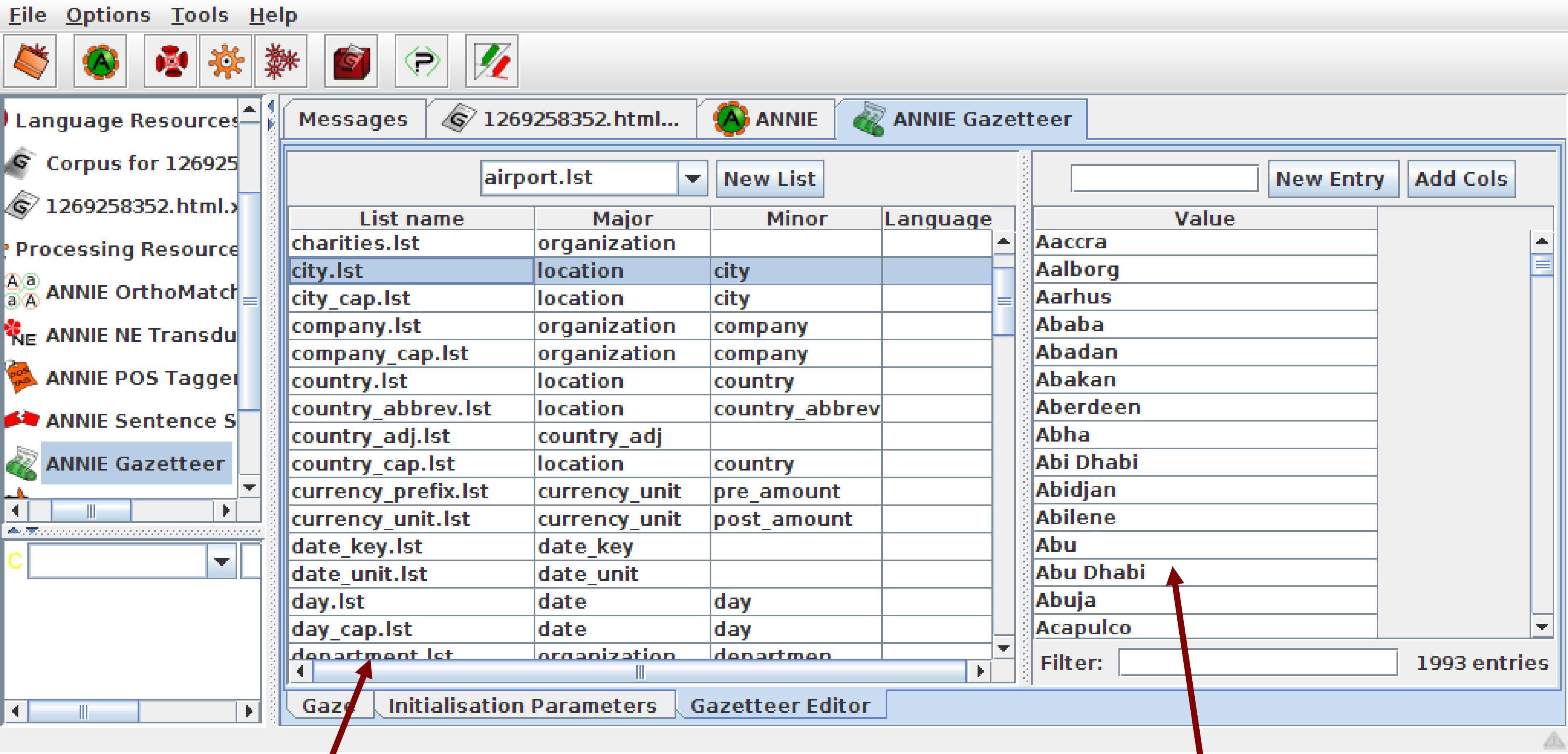
Type	Set	Start	End	Id	Features
Token		0	5	609	{category=NN, kind=word, length=5, orth=upperInitial, root=house, string=H
Token		6	12	611	{affix=s, category=NNS, kind=word, length=6, orth=lowercase, root=price, s
Token		13	15	613	{category=IN, kind=word, length=2, orth=lowercase, root=in, string=in}
Token		16	23	615	{category=NNP, kind=word, length=7, orth=upperInitial, root=england, strin
Token		24	27	617	{category=CC, kind=word, length=3, orth=lowercase, root=and, string=and}
Token		28	33	619	{category=NNP, kind=word, length=5, orth=upperInitial, root=wales, string=
Token		34	38	621	{affix=ed, category=VBD, kind=word, length=4, orth=lowercase, root=be, st
Token		39	41	623	{category=CD, kind=number, length=2, root=10, string=10}
Token		41	42	624	{category=., kind=punctuation, length=1, root=., string=.}
Token		42	43	625	{category=CD, kind=number, length=1, root=8, string=8}
Token		44	47	627	{category=IN, kind=word, length=3, orth=lowercase, root=per, string=per}
Token		48	52	629	{category=NN, kind=word, length=4, orth=lowercase, root=cent, string=cen
Token		53	59	631	{category=JJR, kind=word, length=6, orth=lowercase, root=higher, string=hi
Token		60	62	633	{category=IN, kind=word, length=2, orth=lowercase, root=in, string=in}

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

Gazetteer editor



airport.lst

List name	Major	Minor	Language
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	
day_cap.lst	date	day	
department.lst	organization	departmen	

Value
Aaccra
Aalborg
Aarhus
Ababa
Abadan
Abakan
Aberdeen
Abha
Abi Dhabi
Abidjan
Abilene
Abu
Abu Dhabi
Abuja
Acapulco

Filter: 1993 entries

definition file entries

entries for selected list

ANNIE gazetteer

- Create an ANNIE Gazetteer PR, but change the listsURL to gazetteer/lists.def in the hands-on materials for this module. Double-click on it to open.
- 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 lists and entries can be added by typing in the “New list” or “New entry” box respectively

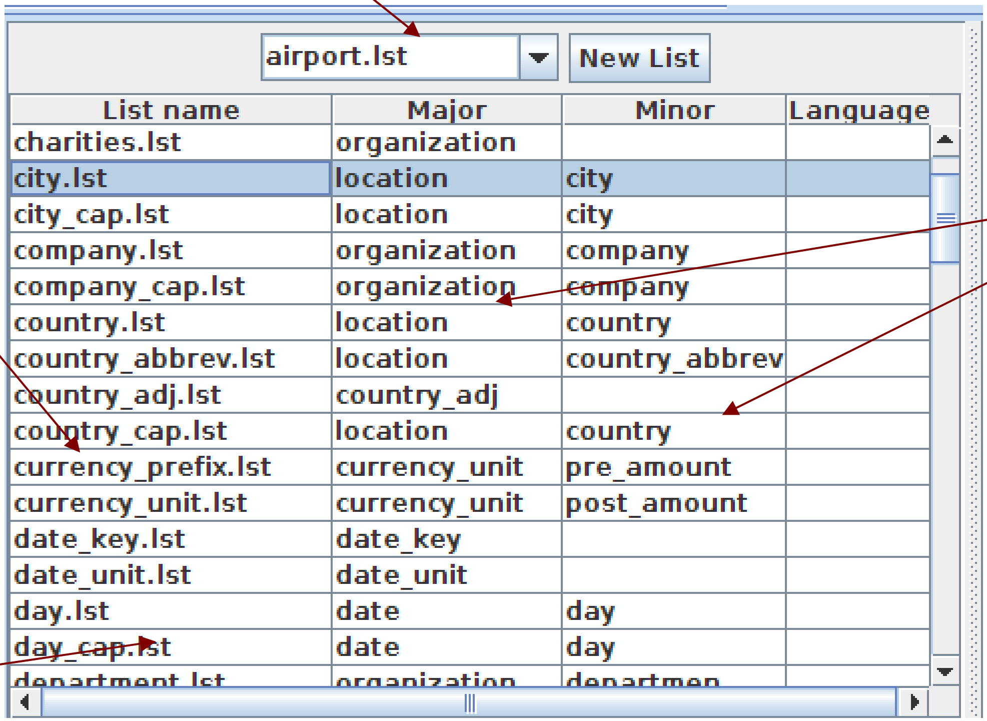
Modifying the definition file

add a new list

edit an existing list name by typing here

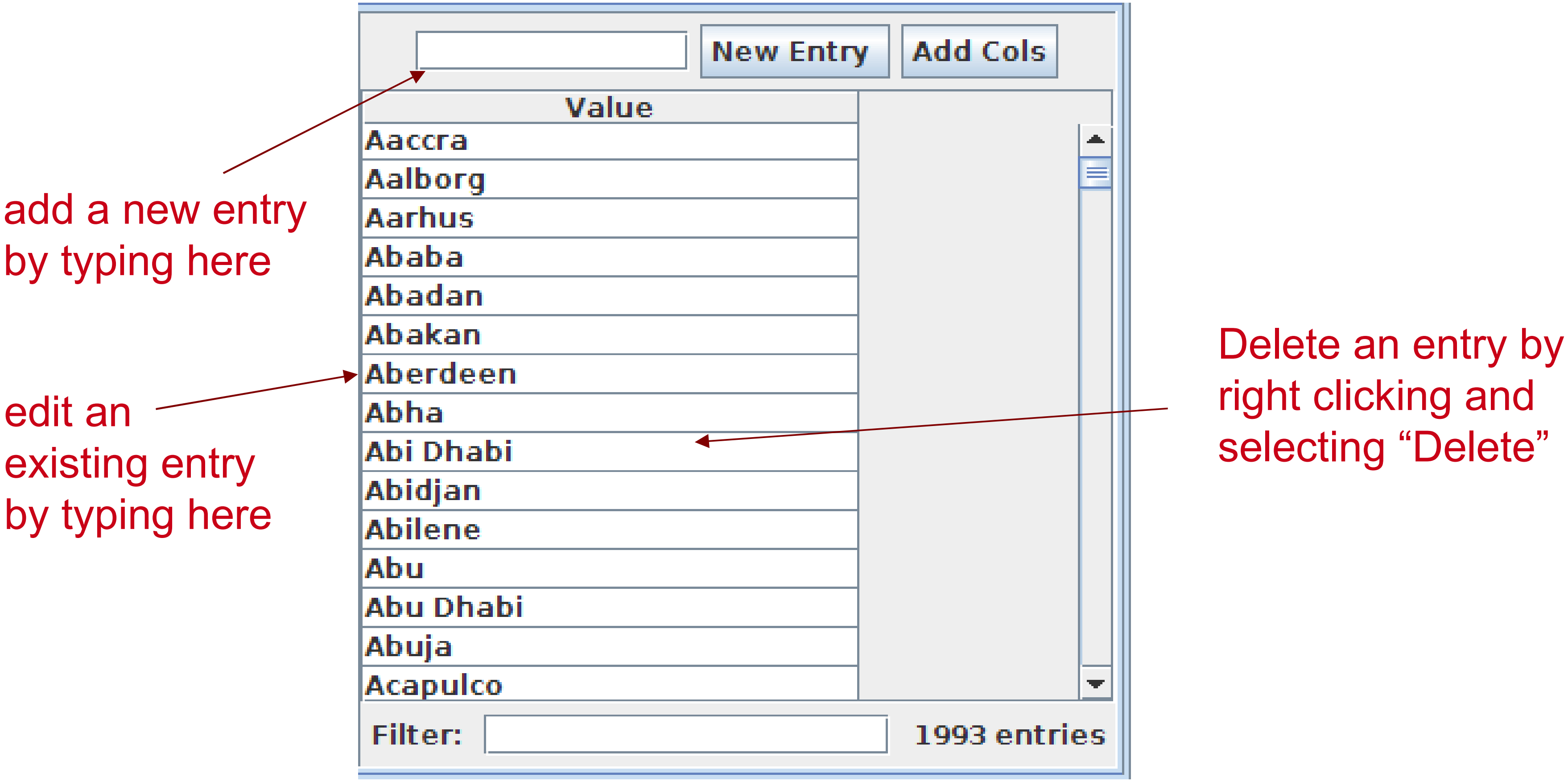
edit the major and minor Types by typing here

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



List name	Major	Minor	Language
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	
day_cap.lst	date	day	
department.lst	organization	departmen	

Modifying a list



Value
Aaccra
Aalborg
Aarhus
Ababa
Abadan
Abakan
Aberdeen
Abha
Abi Dhabi
Abidjan
Abilene
Abu
Abu Dhabi
Abuja
Acapulco

1993 entries

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 left, and select “Save and reinitialise”. (You can select “Save as...” to save your changes in a different location.)
- 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
- You need to reinitialise the gazetteer in GATE (in order to re-load the modified files) 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"

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
- Each NE transducer consists of one or more grammars written in the JAPE language, which Module 3 will cover in detail tomorrow

ANNIE NE Transducer

- Create an ANNIE NE Transducer PR with the default parameters
- 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]

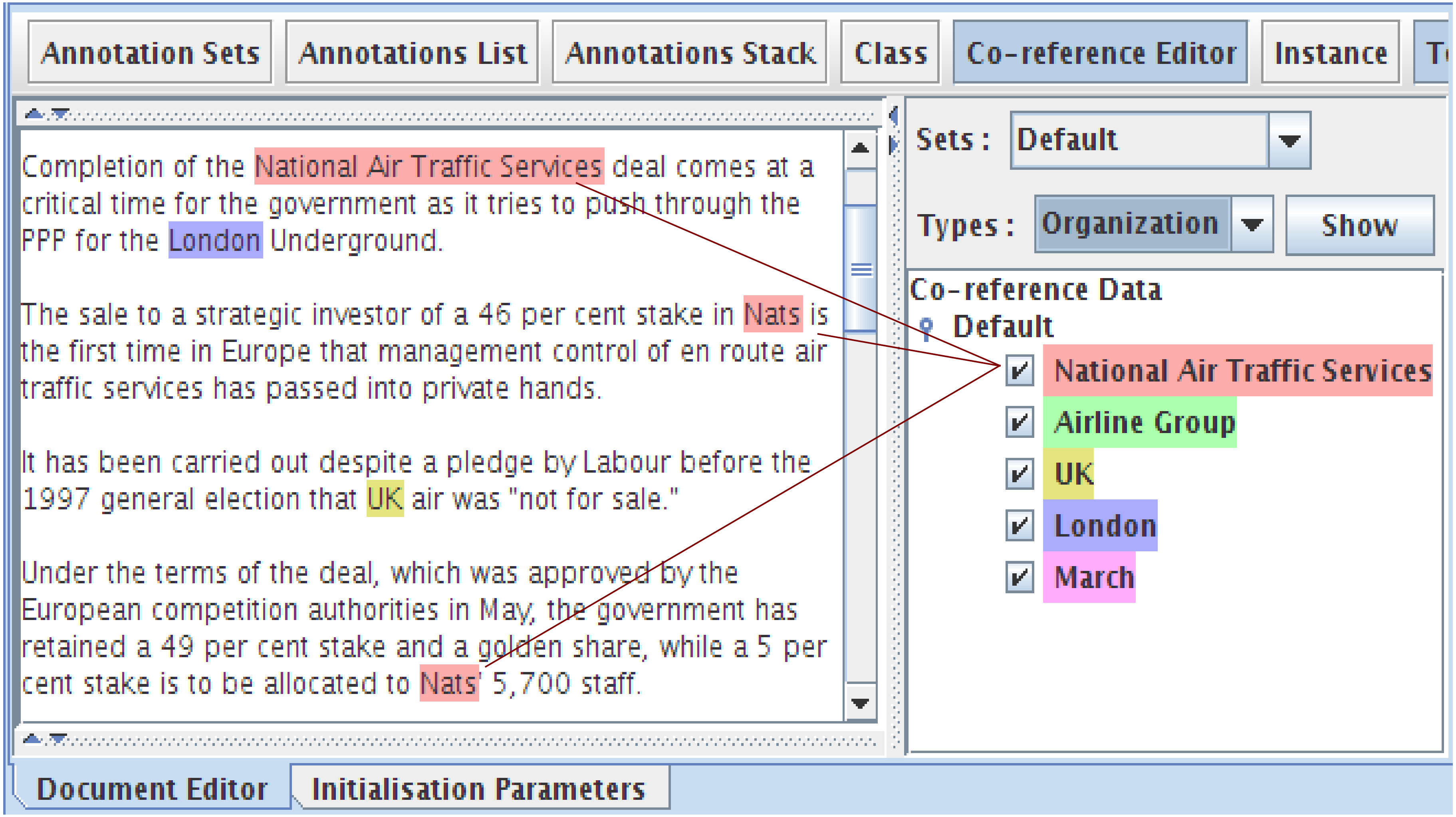
Orthomatcher PR

- 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 co-reference PR is also available in the ANNIE plugin

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



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.

Sets : Default

Types : Organization

Co-reference Data

Default

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

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!
- Deselect all items in this list, then select a type from the “Type” combo box and click “Show” to view all coreferences of a particular annotation type (note that some types may not have coreferences)

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
- Load ANNIE as before, but this time select “Without defaults”
- For each PR, select the default option, except for the gazetteer, where you should select your saved gazetteer index file (gazetteer/lists.def) from the hands-on materials

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 are not plugins, strictly speaking, as they do not provide new kinds of PRs
- 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

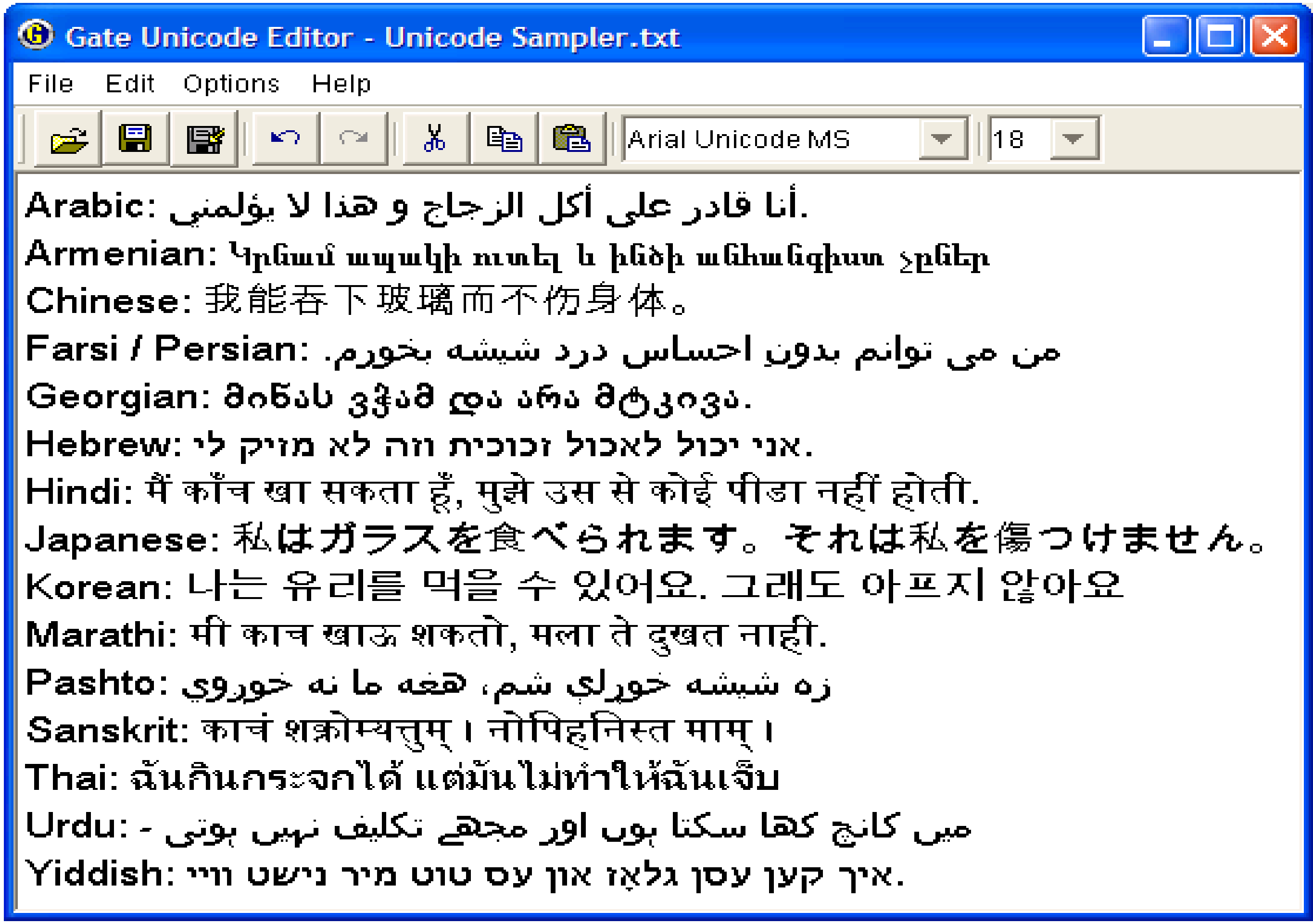
- The following PRs are largely language-independent:
  - Unicode tokeniser
  - Sentence splitter
  - Gazetteer PR (but do localise the lists!)
  - Orthomatcher (depending on the nature of the language)
- Other PRs will need to be adapted (e.g. JAPE transducer) or replaced with a language-specific version (e.g. POS tagger)
- This topic is covered in more detail in Module 10 (Advanced IE, in Track 3)

Useful Multilingual PRs

- Stemmer_Snowball plugin
  - Stemmer PRs with models for Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish (set the language init parameter with one of these in lower case)
  - Run a Tokeniser first (Unicode one is best)
- Tagger_Framework
  - GenericTagger PR which can run external tools such as the TreeTagger

Displaying multilingual data

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

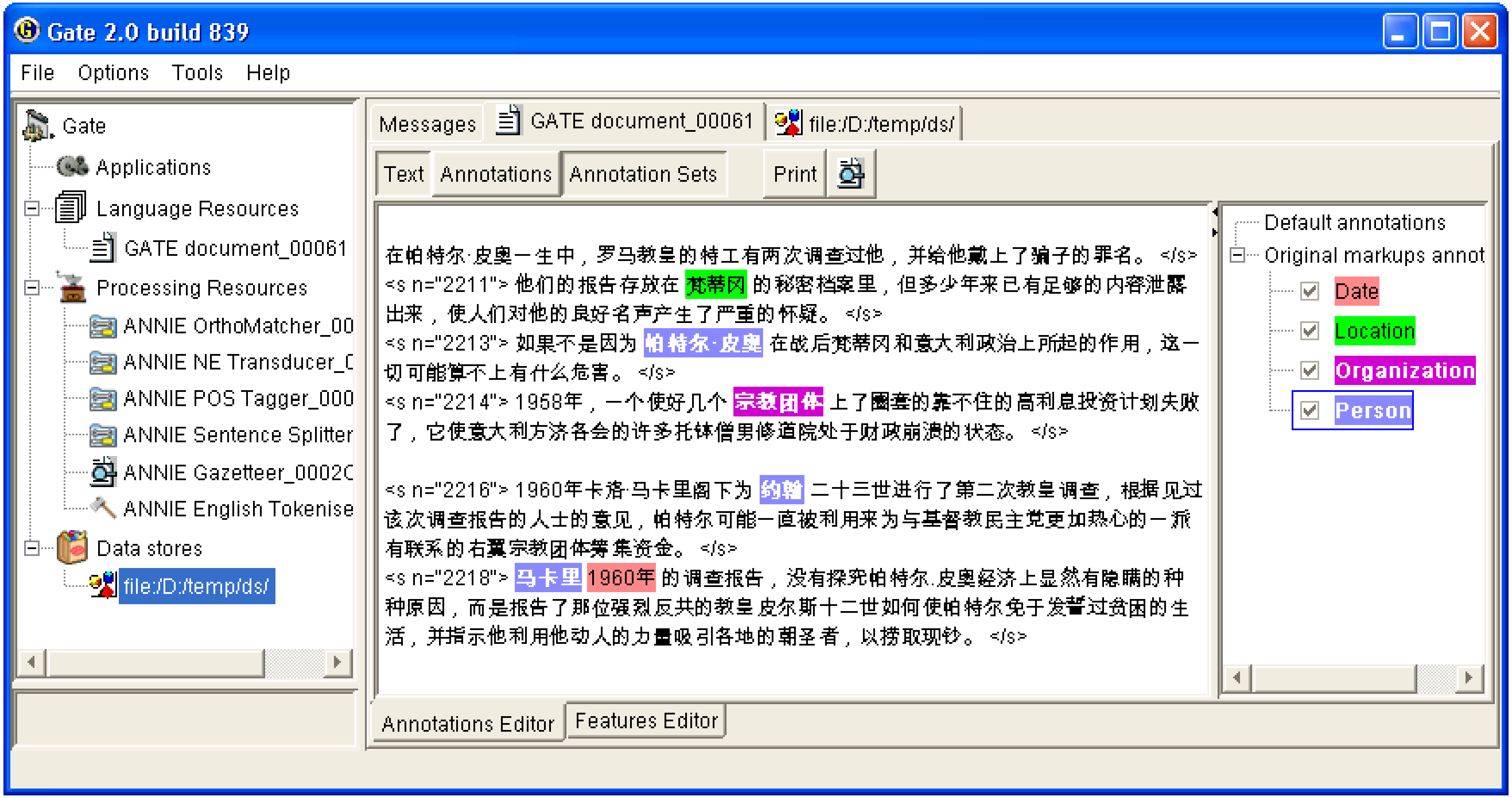


Arial Unicode MS 18

Arabic: أنا قادر على أكل الزجاج و هذا لا يؤلمني.
Armenian: Կրնամ ապակի ուտել և ինծի անհանգիստ չըներ
Chinese: 我能吞下玻璃而不伤身体。
Farsi / Persian: من می توانم بدونِ احساس درد شيشه بخورم.
Georgian: მინას ვჭამ და არა მტკივა.
Hebrew: אני יכול לאכול זכוכית וזה לא מזיק לי.
Hindi: मैं काँच खा सकता हूँ, मुझे उस से कोई पीडा नहीं होती.
Japanese: 私はガラスを食べられます。それは私を傷つけません。
Korean: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요
Marathi: मी काच खाऊ शकतो, मला ते दुखत नाही.
Pashto: زه شيشه خورلي شم، هغه ما نه خوروي
Sanskrit: काचं शक्नोम्यत्तुम् । नोपहिनस्ति माम् ।
Thai: ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ
Urdu: میں کانچ کھا سکتا ہوں اور مجھے تکلیف نہیں ہوتی -
Yiddish: איך קען עסן גלאָז און עס טוט מיר נישט װײ.

Displaying multilingual data

All visualisation and editing tools use the same facilities



在帕特尔·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"> 如果不是因为帕特尔·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，帕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，没有探究帕特尔.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帕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>

Editing multilingual data

- Java currently provides no special support for text input, but the GATE Unicode Kit (GUK) allows the definition of additional Input Methods (currently 30 IMs for 17 languages)
- Pluggable in other applications (e.g. MPI’s EUDICO)
- Can use virtual keyboard or standard layouts over QWERTY
- IMs defined in plain text files
- Use Options → Input methods to change the keyboard mapping
- **However**, if your OS provides a way to change keyboard mappings on the fly (as most do these days) and covers the language you want to use, you should use that function instead.

Annotation and Evaluation

Topics covered

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

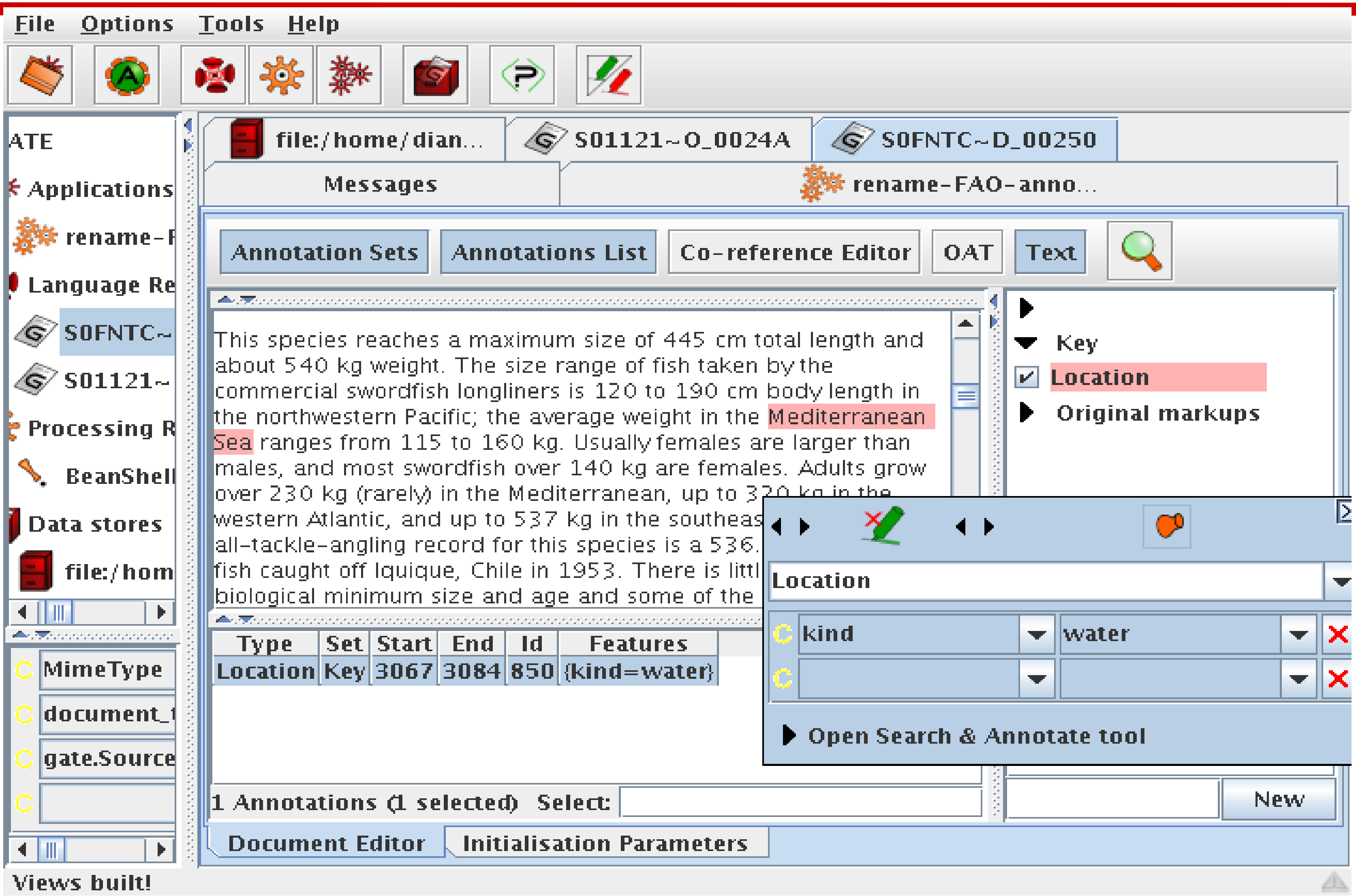
Before you start annotating...

- 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



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 southeas all-tackle-angling record for this species is a 536. fish caught off Iquique, Chile in 1953. There is littl biological minimum size and age and some of the

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

1 Annotations (1 selected) Select:

Location

kind water

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

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

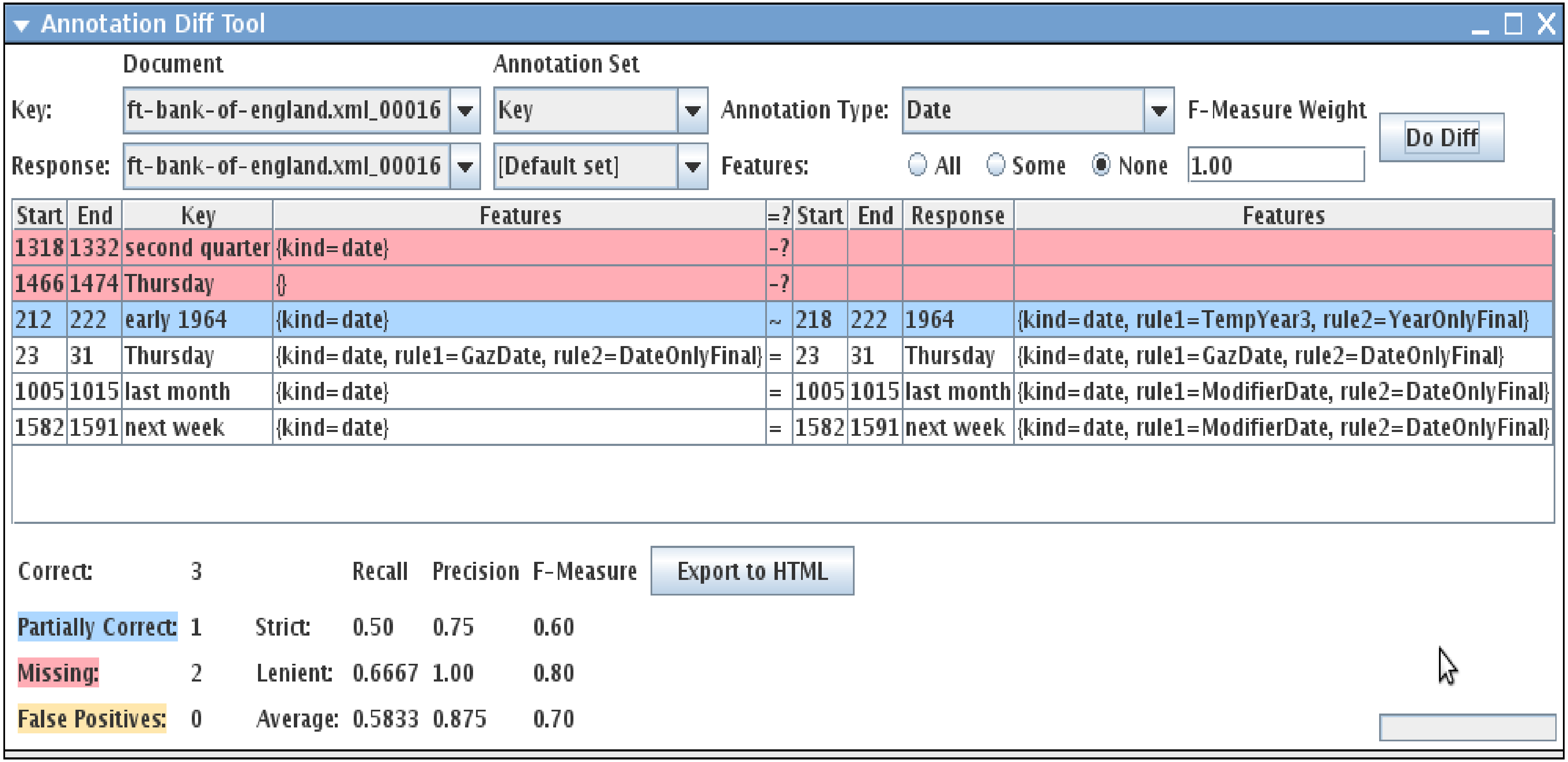
Annotations are like squirrels…

Annotation Diff helps with “spot the difference”

Annotation Diff Exercise

- Open the document in-whitbread-10-aug-2001.xml (after processing with ANNIE), create a new Key annotation set and add some new Person annotations there. Add some incorrect annotations as well as correct ones.
- Open the AnnotationDiff: Tools → Annotation Diff or click the icon (green and red pencils)
- Select the name of the document you annotated
- Key contains the manual annotations (select **Key** annotation set)
- Response contains annotations from ANNIE (select **Default** annotation set)
- Select the **Person** annotation
- Click on “Compare”

Annotation Diff



	Document	Annotation Set		
Key:	ft-bank-of-england.xml_00016	Key	Annotation Type:	Date
Response:	ft-bank-of-england.xml_00016	[Default set]	Features:	○ All ○ Some ● None

F-Measure Weight: 1.00

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

			Recall	Precision	F-Measure
Correct:	3				
Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

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

Gold Standard (IE)	**Gold Standard (IR)**	**Inter-annotator Agreement**
Correct	True Positive	Match
Missing	False Negative	Only A (or B)
Spurious	False Positive	Only B (or A)
Partially Correct		Overlap
	True Negative	

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

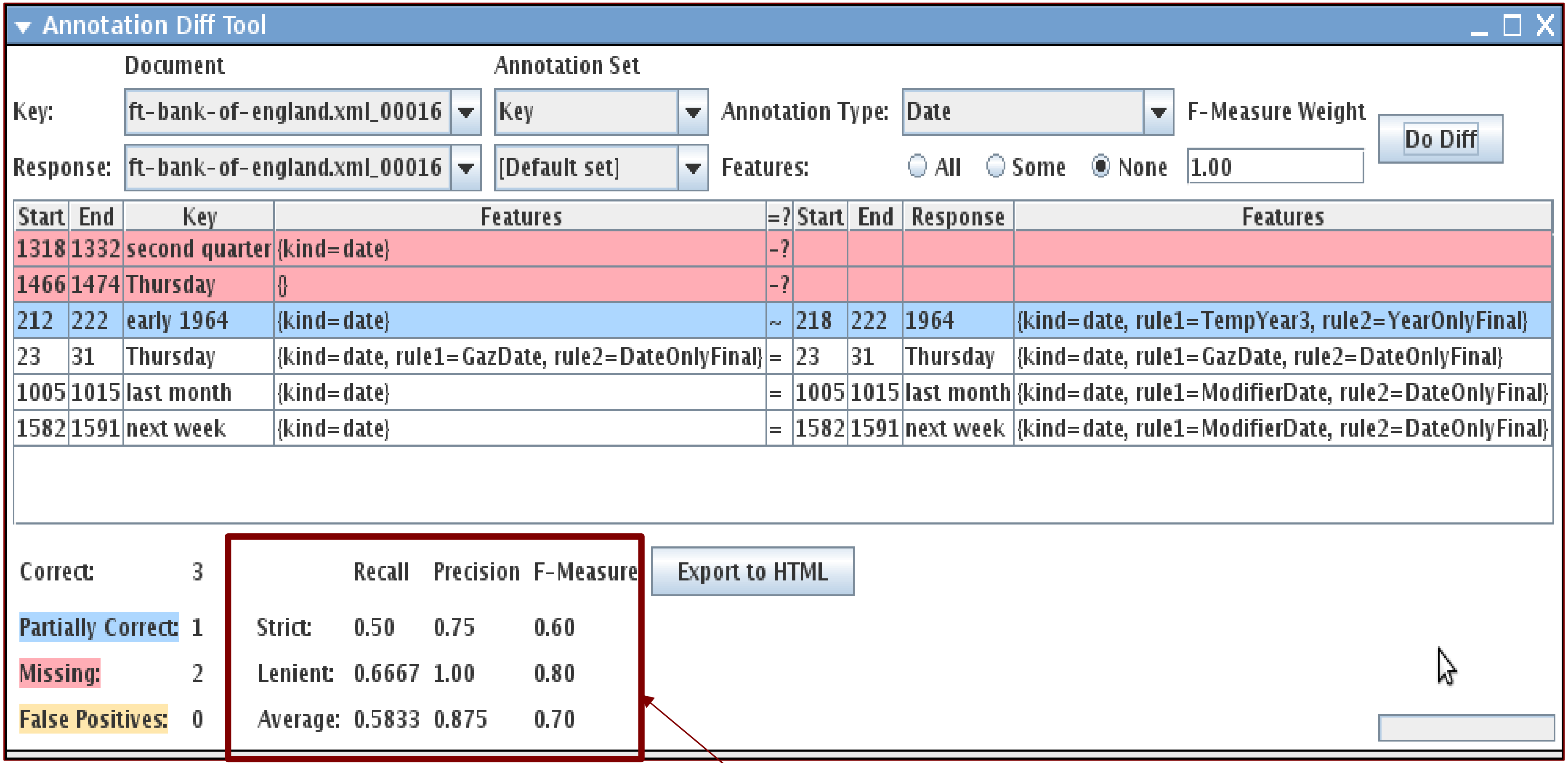
- **Spurious** = things annotated wrongly

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



	Document	Annotation Set
Key:	ft-bank-of-england.xml_00016	Key
Response:	ft-bank-of-england.xml_00016	[Default set]

Annotation Type: Date

Features: ○ All ○ Some ● None

F-Measure Weight: 1.00

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct:	3
Partially Correct:	1
Missing:	2
False Positives:	0

	Recall	Precision	F-Measure
Strict:	0.50	0.75	0.60
Lenient:	0.6667	1.00	0.80
Average:	0.5833	0.875	0.70

scores displayed

Precision

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

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

Recall

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

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

F-Measure

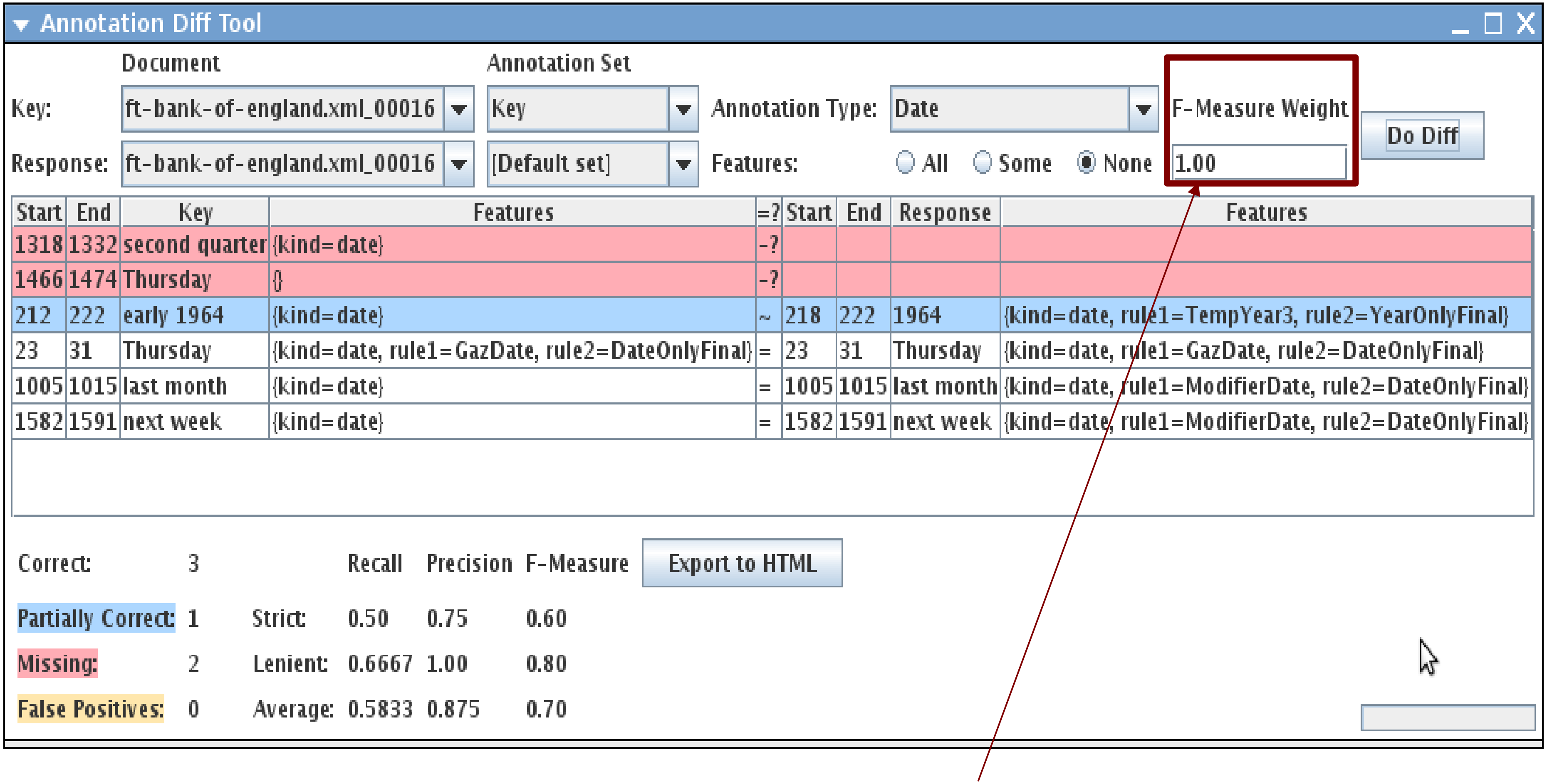
- Precision and recall tend to trade off against one another
  - If you specify your rules precisely to improve precision, you may get a lower recall
- If you make your rules very general, you get good recall, but low precision
- This makes it difficult to compare applications, or to check whether a 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

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

Annotation Diff defaults to F1



	Document	Annotation Set		
Key:	ft-bank-of-england.xml_00016	Key	Annotation Type: Date	F-Measure Weight
Response:	ft-bank-of-england.xml_00016	[Default set]	Features: ○ All ○ Some ● None	1.00

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

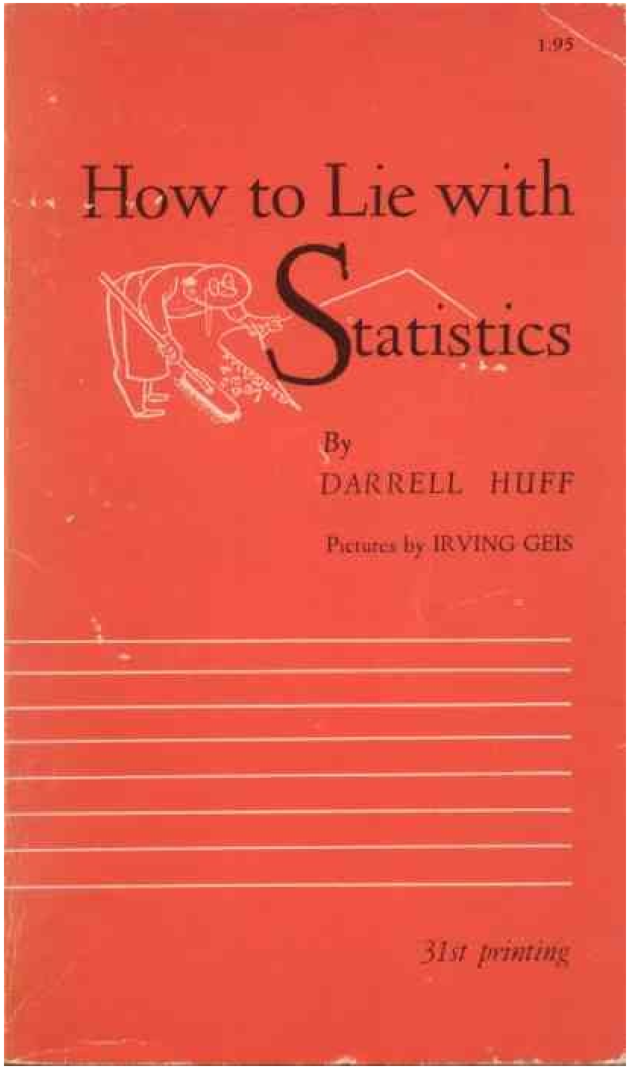
			Recall	Precision	F-Measure
Correct:	3				
Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

F-measure weight set to 1

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

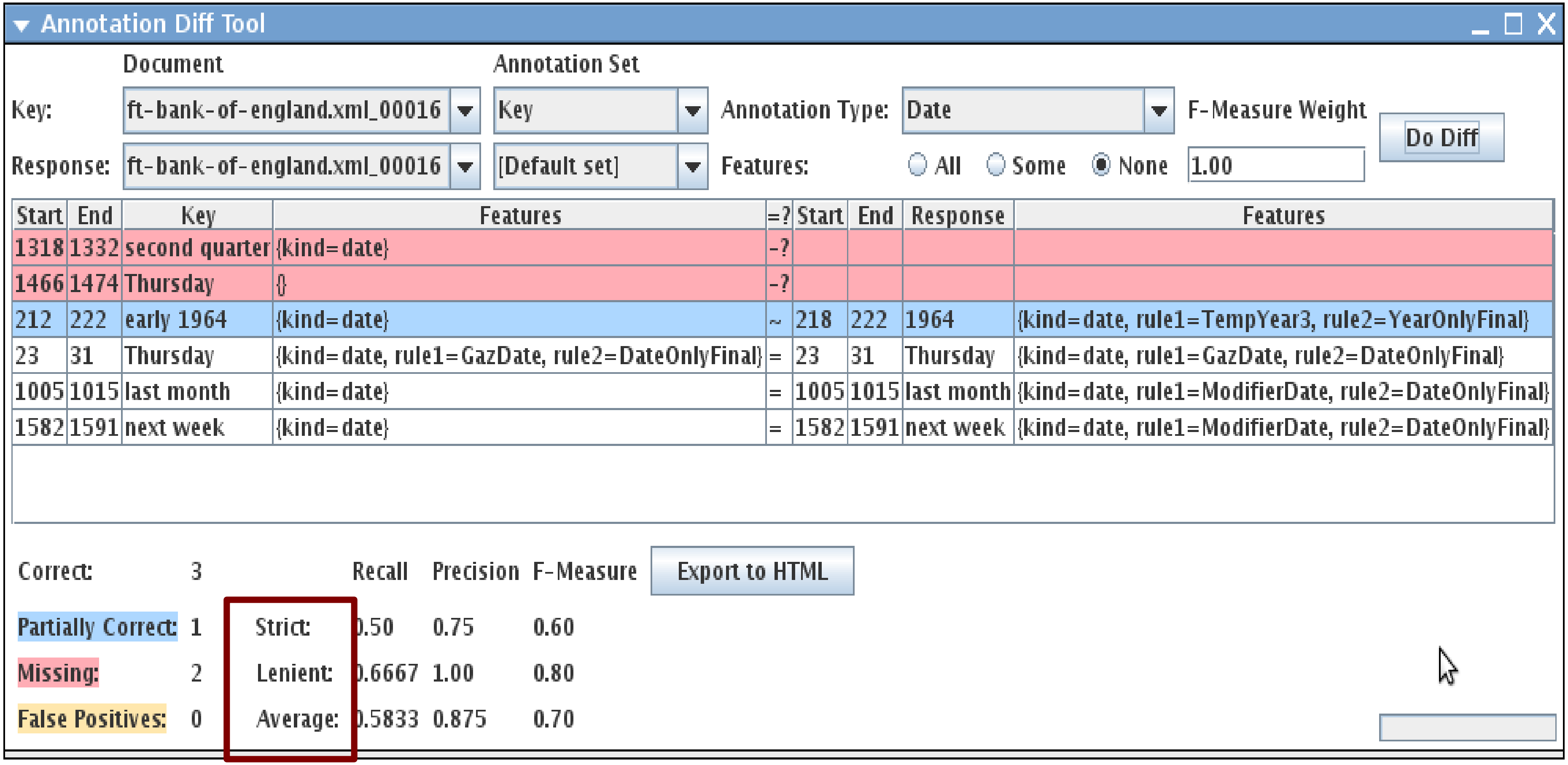
- How we want to measure partially correct annotations may differ, depending on our goal
- In GATE, there are 3 different ways to measure them
- The most usual way is to consider them to be “half right”
- Average: Strict and lenient scores are averaged (this is the same as counting a half weight for 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!

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



Huff (1954) is still the classic text on the subject!

Strict, Lenient and Average



	Document	Annotation Set		
Key:	ft-bank-of-england.xml_00016	Key	Annotation Type:	Date
Response:	ft-bank-of-england.xml_00016	[Default set]	Features:	○ All ○ Some ● None

F-Measure Weight: 1.00

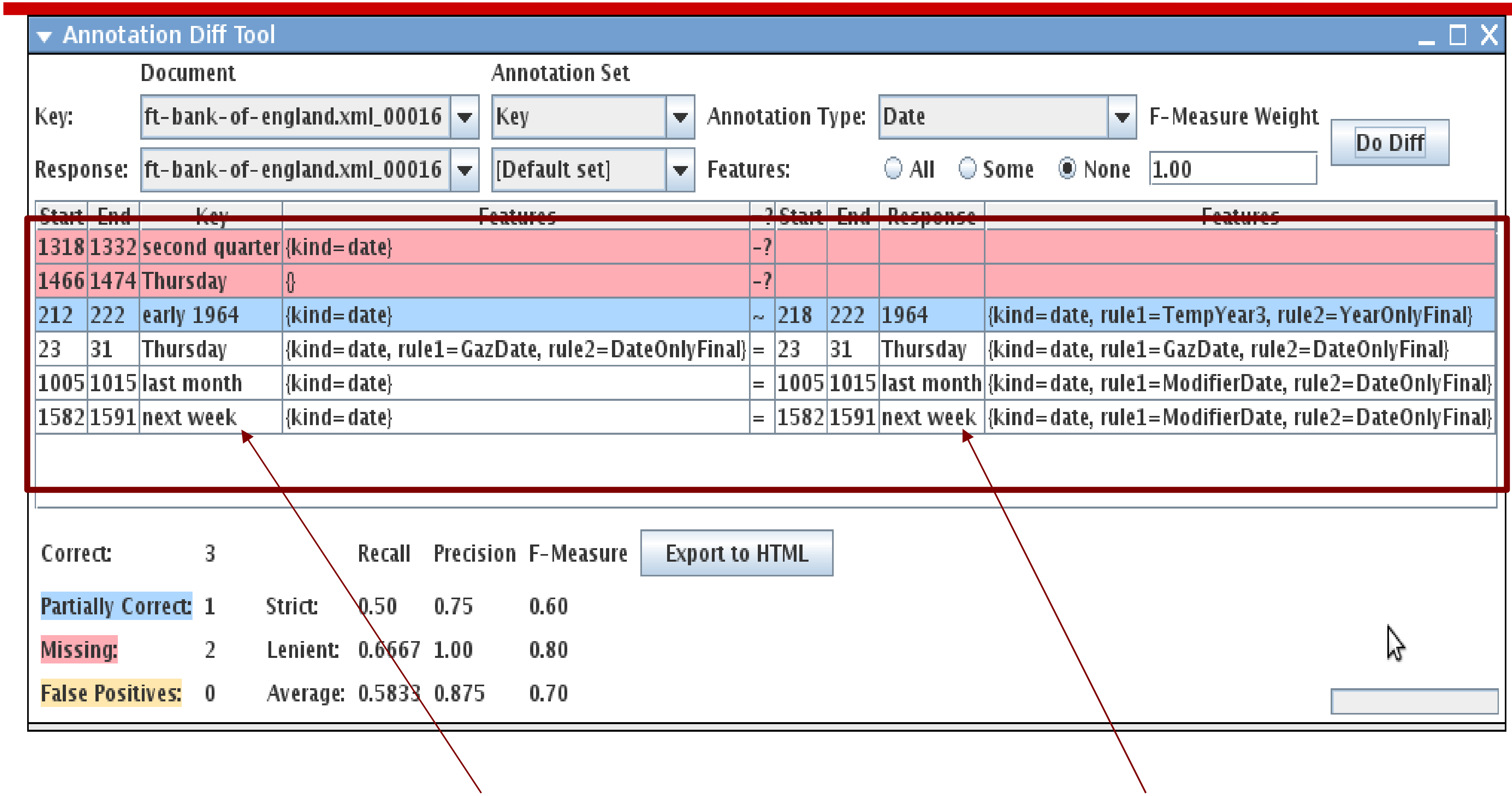
Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

			Recall	Precision	F-Measure
Correct:	3				
Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

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



	Document	Annotation Set
Key:	ft-bank-of-england.xml_00016	Key
Response:	ft-bank-of-england.xml_00016	[Default set]

Annotation Type: Date

Features: ○ All ○ Some ● None

F-Measure Weight 1.00

Start	End	Key	Features	?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

			Recall	Precision	F-Measure
Correct:	3				
Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

Key annotations

Response annotations

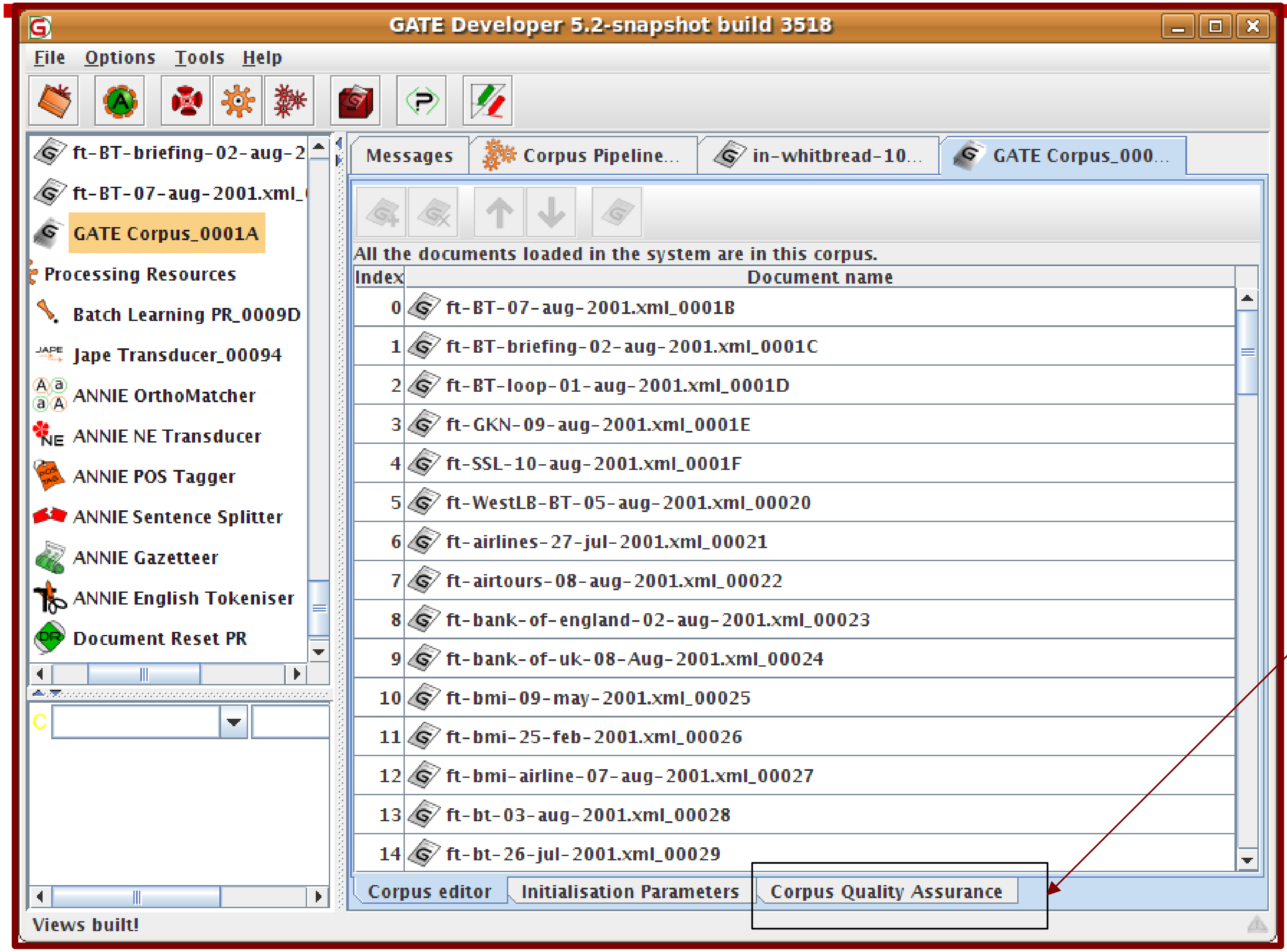
Corpus Quality Assurance

- 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.
- Nine of the documents (ft-*.xml) were manually annotated for you (in the Key AS).
- You annotated one of them (in-whitbread-10-aug-2001.xml).
- Click the Corpus Quality Assurance tab at the bottom of the Display pane.

Try out Corpus Quality Assurance



GATE Developer 5.2-snapshot build 3518

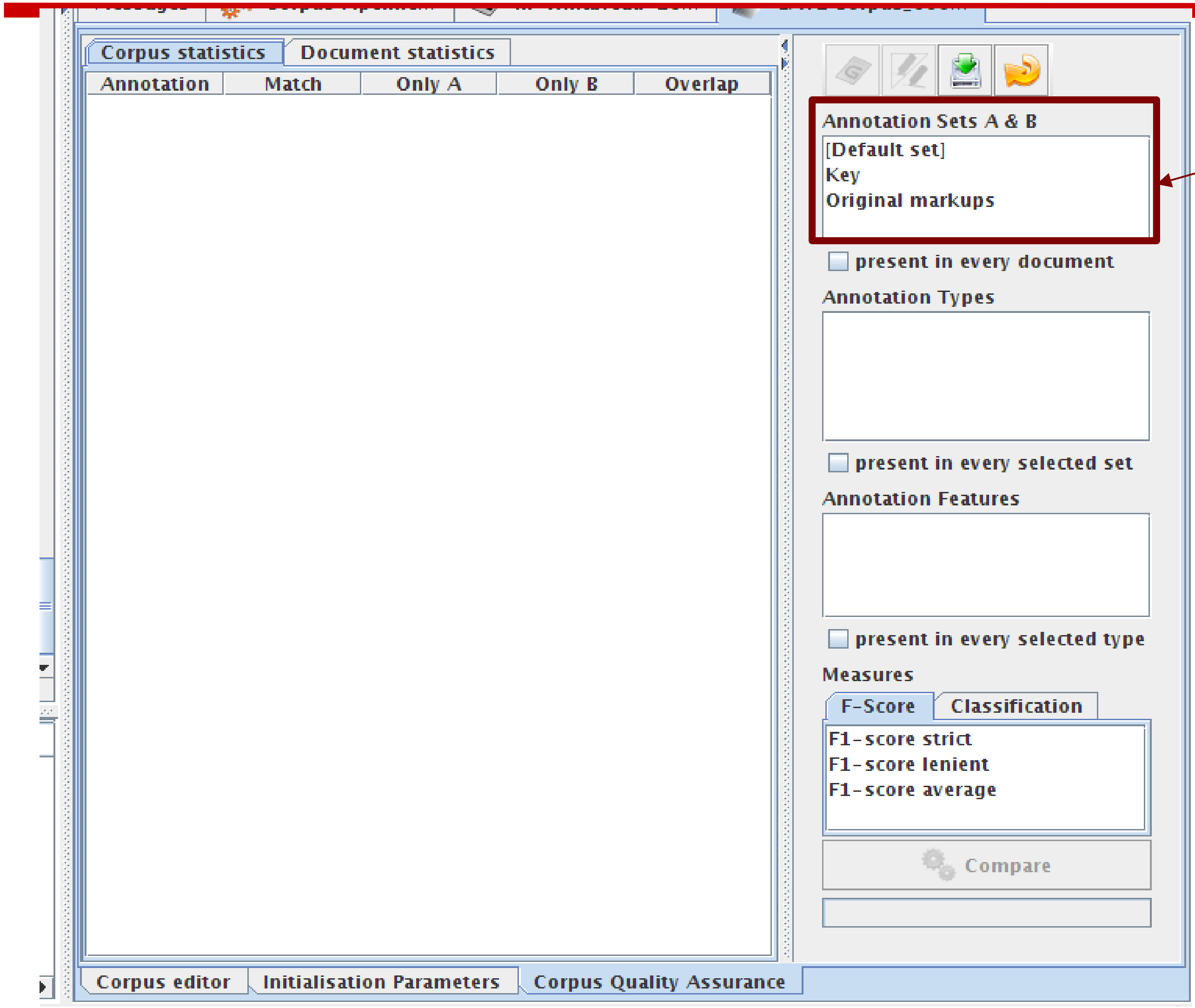
All the documents loaded in the system are in this corpus.

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

Views built!

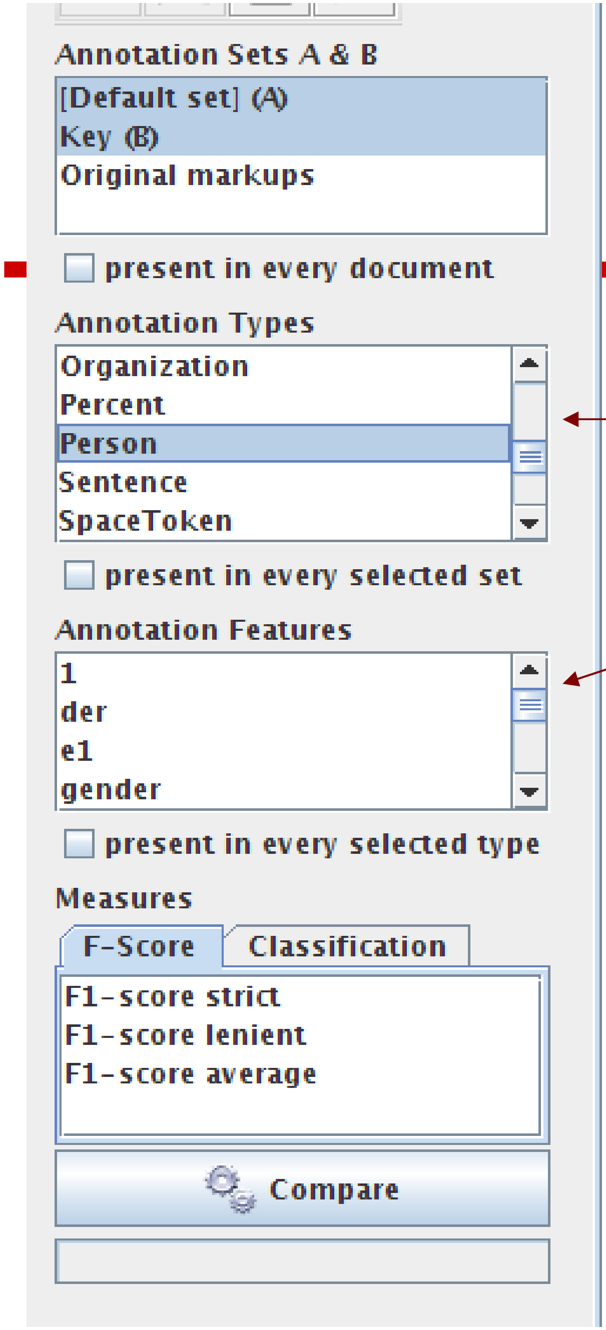
- Click the Corpus Quality Assurance tab at the bottom of the Display pane.

Select Annotation Sets



- Select the annotation sets you wish to compare.
- Click on an annotation set – this will label it set A.
- Now click on another annotation set - this will label it set B

Select Type



Annotation Sets A & B

- [Default set] (A)
- Key (B)
- Original markups

□ present in every document

Annotation Types

- Organization
- Percent
- Person
- Sentence
- SpaceToken

□ present in every selected set

Annotation Features

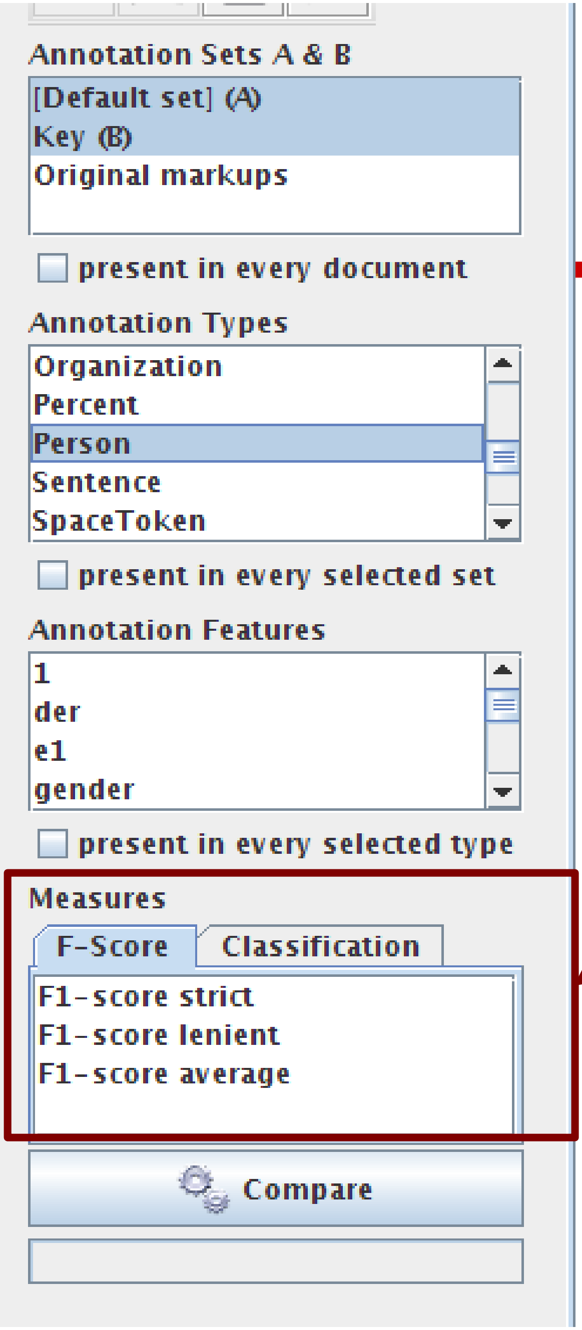
- 1
- der
- e1
- gender

□ present in every selected type

Measures

- F1-score strict
- F1-score lenient
- F1-score average

- Select the annotation type to compare
- Select the features to include (if any)
- You can select as many as you want.



Annotation Sets A & B

[Default set] (A)
Key (B)
Original markups

□ present in every document

Annotation Types

Organization
Percent
Person
Sentence
SpaceToken

□ present in every selected set

Annotation Features

1
der
e1
gender

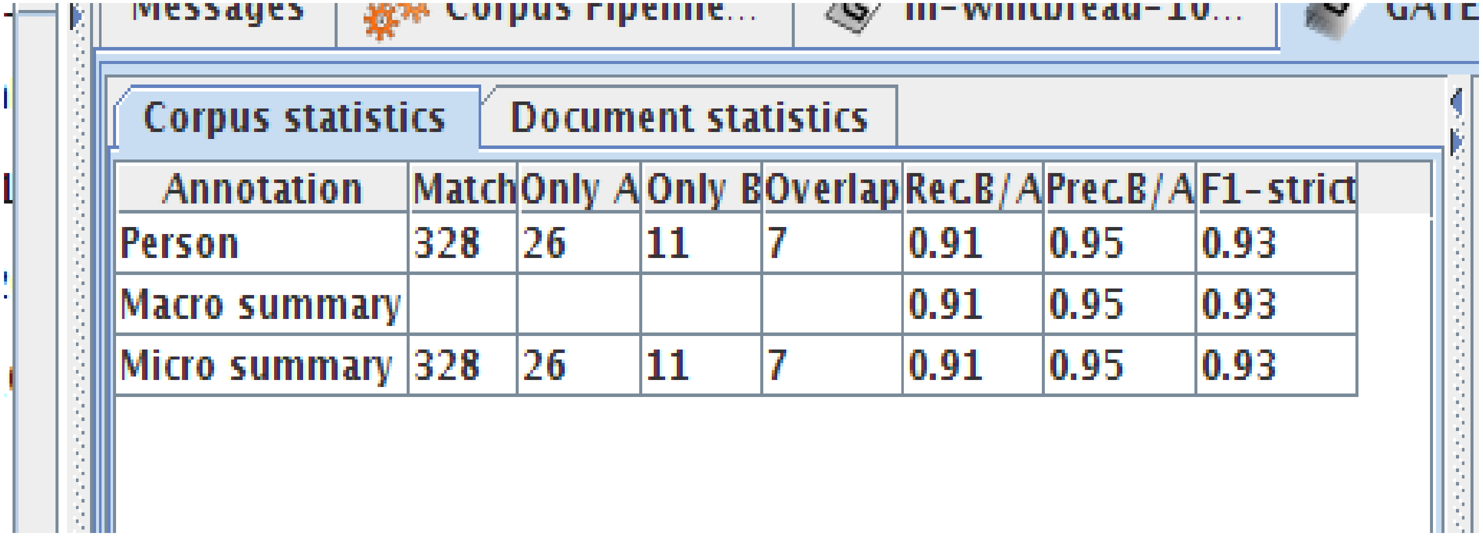
□ present in every selected type

Measures

F1-score strict
F1-score lenient
F1-score average

- In the “Measures” box, select the kind of F score you want “Strict, Lenient, Average” or any combination of them.
- Select Compare

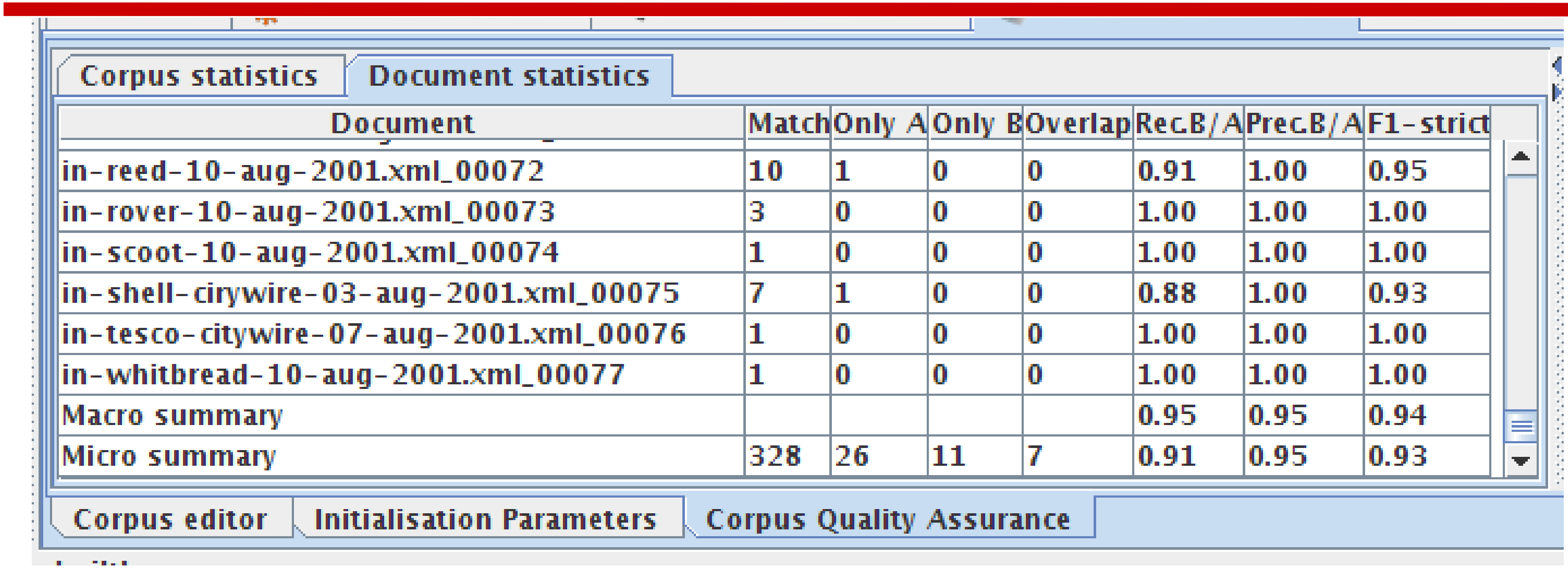
Corpus Statistics Tab



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



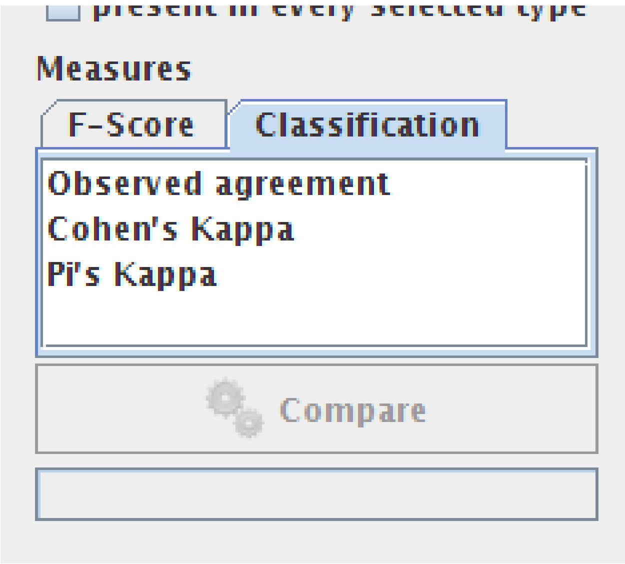
Document	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
in-reed-10-aug-2001.xml_00072	10	1	0	0	0.91	1.00	0.95
in-rover-10-aug-2001.xml_00073	3	0	0	0	1.00	1.00	1.00
in-scoot-10-aug-2001.xml_00074	1	0	0	0	1.00	1.00	1.00
in-shell-cirywire-03-aug-2001.xml_00075	7	1	0	0	0.88	1.00	0.93
in-tesco-citywire-07-aug-2001.xml_00076	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



present in every selected type

Measures

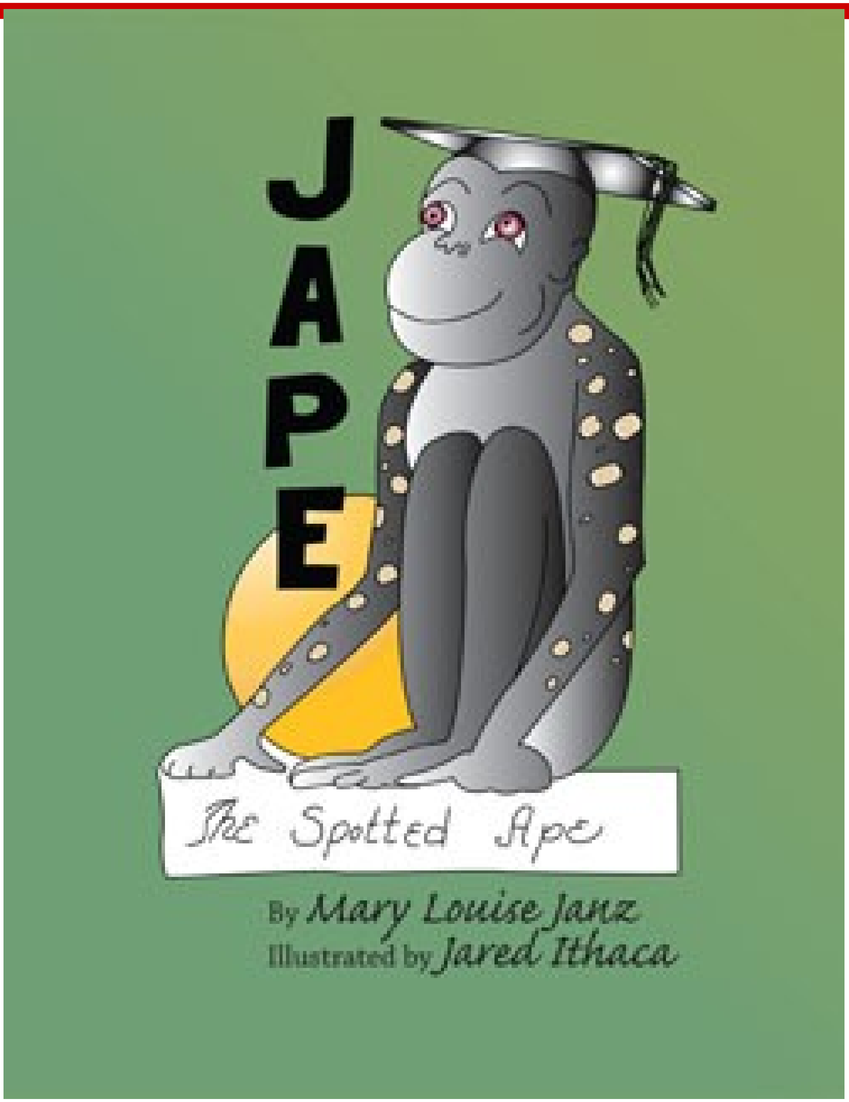
Observed agreement
Cohen's Kappa
Pi's Kappa

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

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 JAPE



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 exercise: 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 Setence annotations.