Module 1 Session 2

Introduction to Information Extraction (IE) and ANNIE

About this tutorial

This tutorial comprises the following topics:

- Introduction to IE
- ANNIE
- Evaluation and Corpus Quality Assurance

In the next session, you’ll learn how to use JAPE, the pattern matching language that is used in ANNIE and useful for many IE tasks.

Information Extraction

- Find relevant parts of texts: e.g. mentions of persons, dates
- Input: unstructured text
- Output: structured information (XML file, Database table, RDF)
- IE is often hard:
  - Has to deal with the ambiguity of language
  - Requires knowledge about the language
  - Requires knowledge about the world
- Information could be e.g.:
  - Persons, Locations, .. mentioned: **Named Entity Recognition**
  - Relations between entities
  - Events

Named Entity Recognition

- Named Entity
  - some specific individual named thing in the world
  - often extends to other pieces of information like dates
- Different **types** of NEs: *Person*, *Organization* (companies, government organizations, committees, universities, etc), *Location* (cities, countries, rivers, etc. bridge? mall?)
- Various other types are frequently added, depending on the task, e.g. newspapers, ships, species, monetary amounts, percentages.
- Need to **find** and **classify** the mentions

NER Importance

- NER provides a foundation from which to build more complex IE systems.
  Once we have found NEs we can:
- Find co-referring mentions: “Dr Smith”, “John Smith”, “John”, “he”
- Find relations: “He (Dr Smith) is CEO of XY”
- Link to Ontologies/Knowledge Bases: “Athens, Georgia” vs “Athens, Greece” (*Named Entity Disambiguation*)
- ...

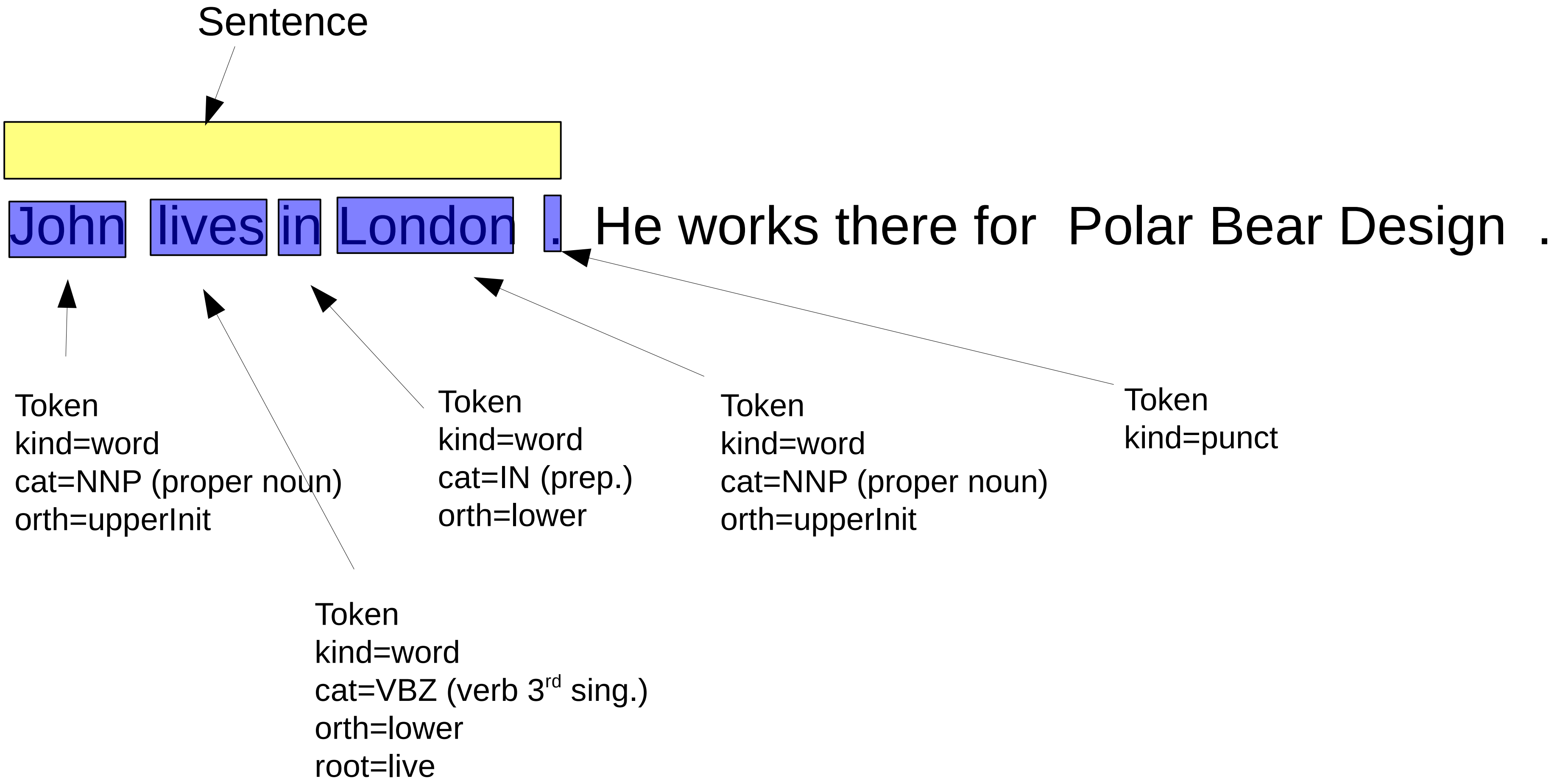
Typical NER pipeline

- Pre-processing
  - Tokenisation: characters → tokens/words
  - Sentence splitting
  - POS tagging: nouns, verbs, adjectives, ...
  - Morphological analysis: word root/lemma
- Find entity mentions: Persons, Locations, …
- Type disambiguation (Person or Location or ..?)
- Coreferencing: (different) mentions of the same entity

Example of IE

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

Preprocessing



Sentence

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

Token
kind=word
cat=NNP (proper noun)
orth=upperInit

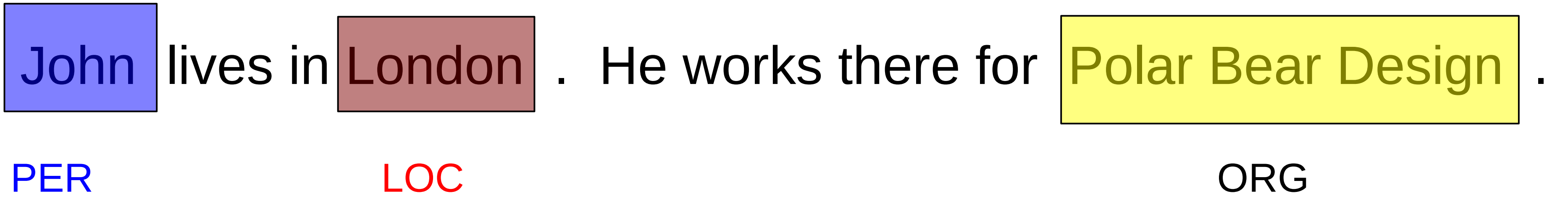
Token
kind=word
cat=IN (prep.)
orth=lower

Token
kind=word
cat=NNP (proper noun)
orth=upperInit

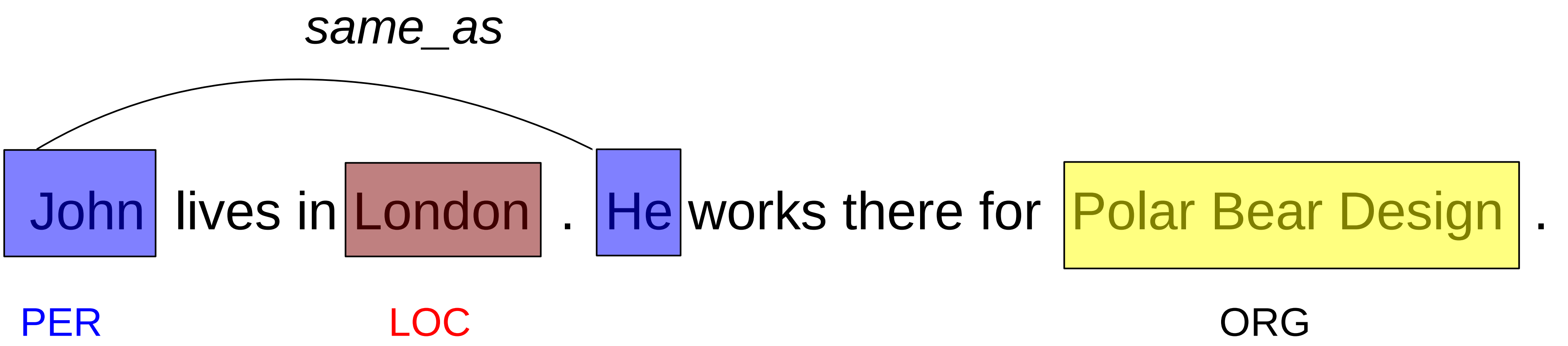
Token
kind=punct

Token
kind=word
cat=VBZ (verb 3rd sing.)
orth=lower
root=live

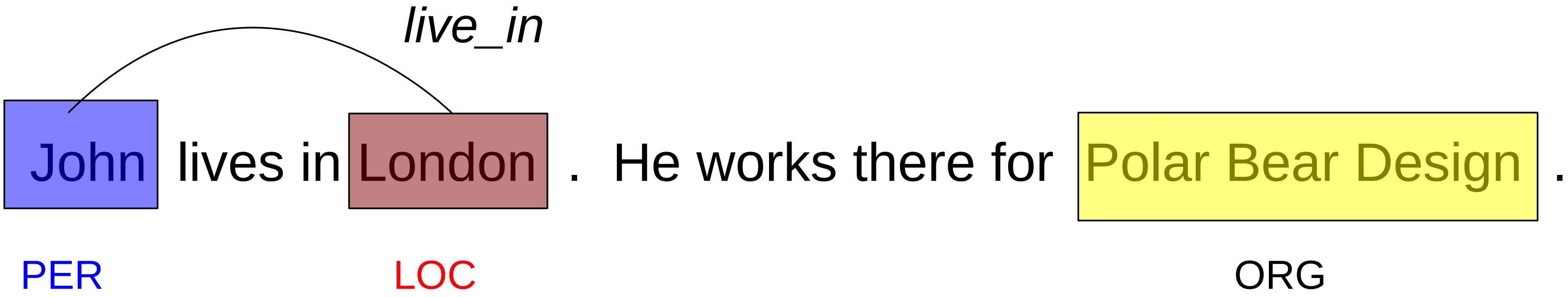
Basic NE Recognition



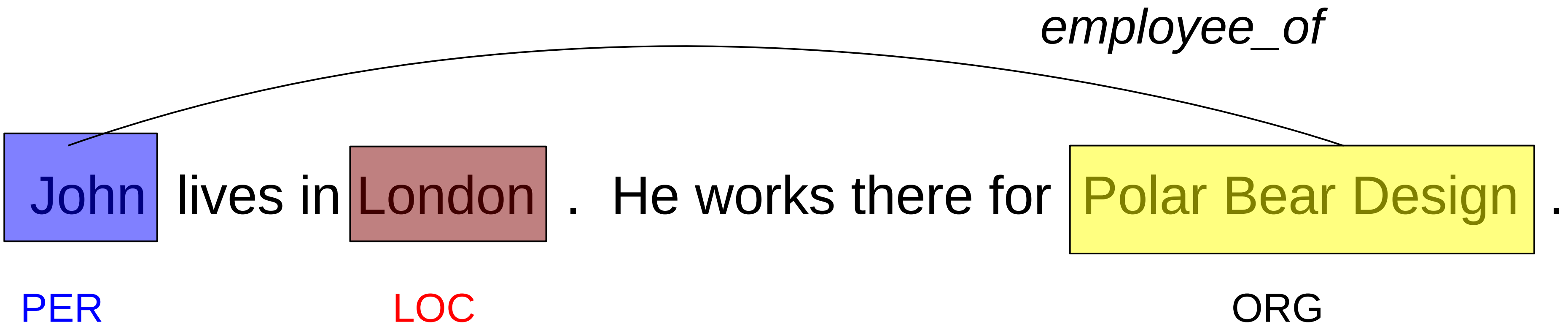
Co-reference



Relations


live_in
John
lives in
London
.
He works there for
Polar Bear Design
.
PER
LOC
ORG

Relations (2)



employee_of

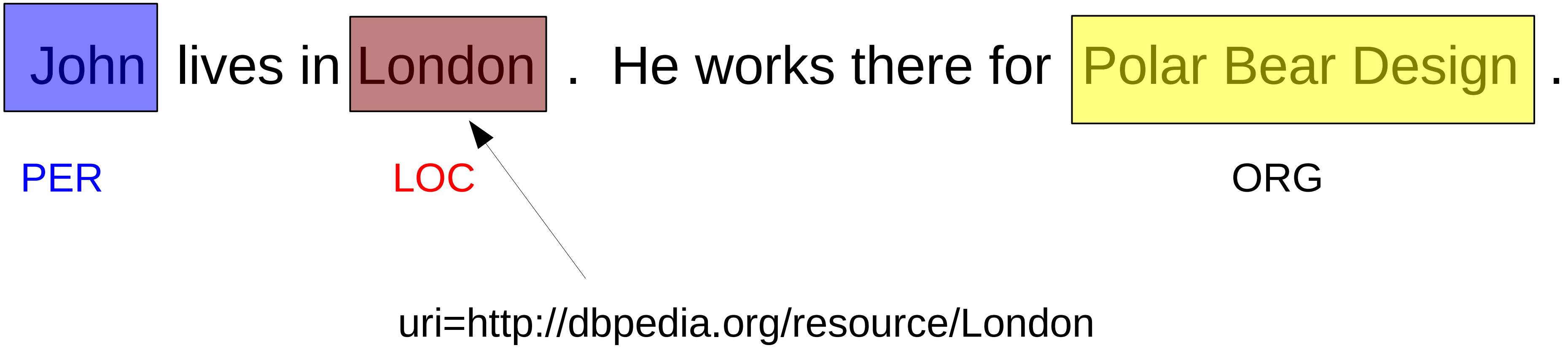
John lives in London . He works there for Polar Bear Design .

PER

LOC

ORG

[Semantic Annotation / Linking]



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

PER

LOC

ORG

uri=http://dbpedia.org/resource/London

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 your hands-on directory `annie-hands-on/news-texts`
- Things for you to try yourself are in **red**.

➢ **Start GATE on your computer now (if you haven't already)**

A Nearly New Information Extraction [System]

- ANNIE is a ready made collection of processing resources (PRs) that performs (parts of) IE on text.
- ANNIE was “nearly new” because
  - It was based on an existing IE system, LaSIE
  - Being >10 years old now, it's not really new any more
- ANNIE performs competitively “out of the box”
- ANNIE can be used as the starting point for your own solutions

ANNIE Components

- The ANNIE application contains a set of core PRs:
  - Tokeniser
  - Sentence Splitter
  - POS tagger
  - Gazetteers
  - Named entity tagger (JAPE transducer)
  - Orthomatcher (orthographic NE coreference)
- Other PRs available from the plugin (not in the application):
  - Pronominal Coreferencer
  - Additional PR s in the Tools plugin,
    e.g. Morphological Analyser, VP Chunker

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 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**
- **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)

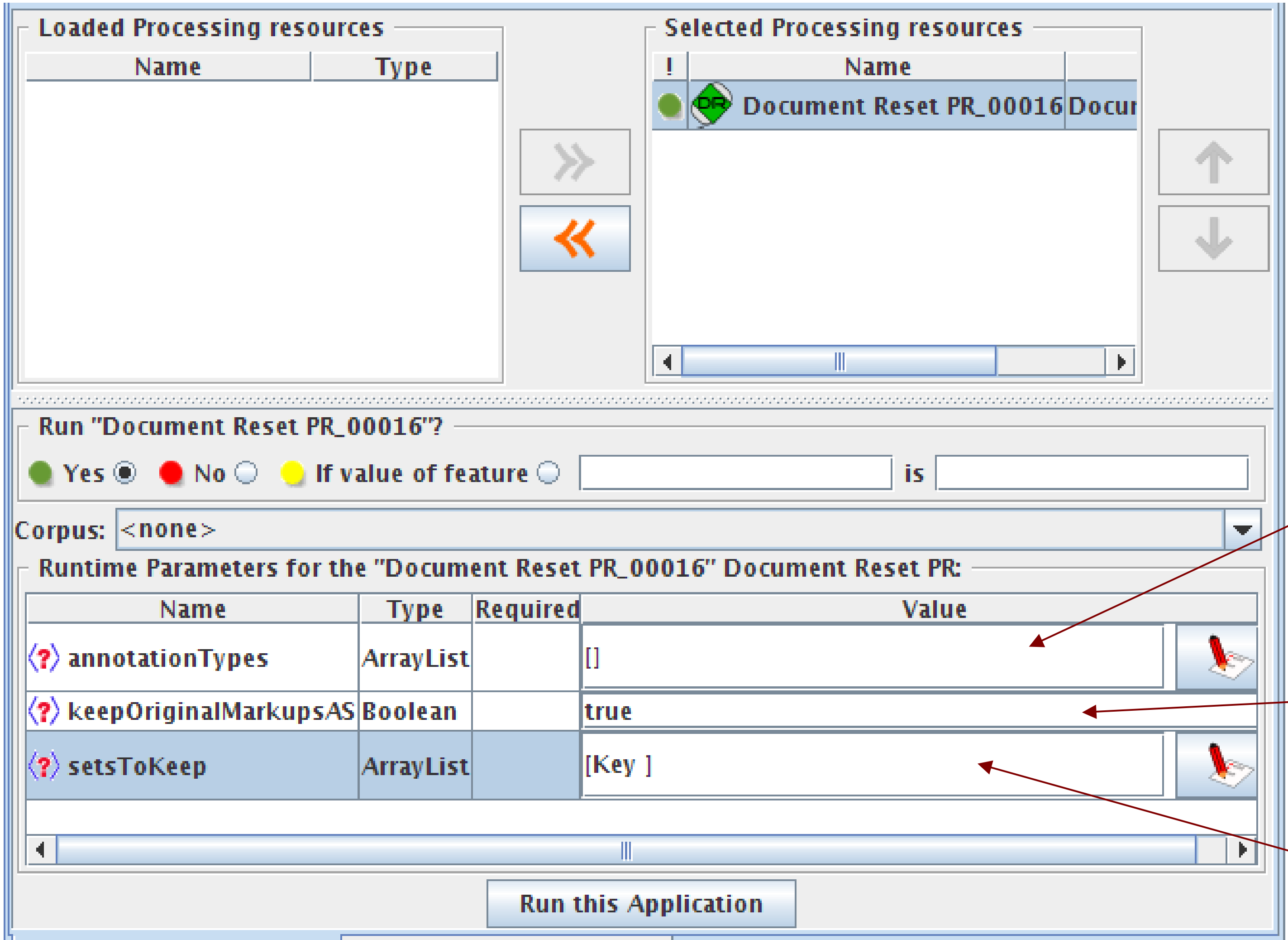
ANNIE Processing Resources

- **View the ANNIE application**
  (Double Click the Icon near the name)
- Each PR in the ANNIE pipeline creates some new annotations, or modifies existing ones
- Document Reset: removes any existing annotations
- Tokeniser: create Token, SpaceToken annotations
- Gazetteer: create Lookup annotations
- Sentence Splitter: create Sentence, Split annotations
- POS tagger: adds category feature to Token annotations
- NE transducer: create 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
- It keeps the “Key” set, by default
- You can configure it to keep any other annotation sets you want, or to remove particular annotation types only
- In more complex applications it is a good idea to:
  - At the start, delete only the annotations created
  - At the end, delete any temporary annotations

Document Reset Parameters



Loaded Processing resources

Name	Type

Selected Processing resources

!	Name	
●	Document Reset PR_00016	Docu

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
annotationTypes	ArrayList		[]
keepOriginalMarkupsAS	Boolean		true
setsToKeep	ArrayList		[Key]

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

Keep Original Markups set

Keep Key set

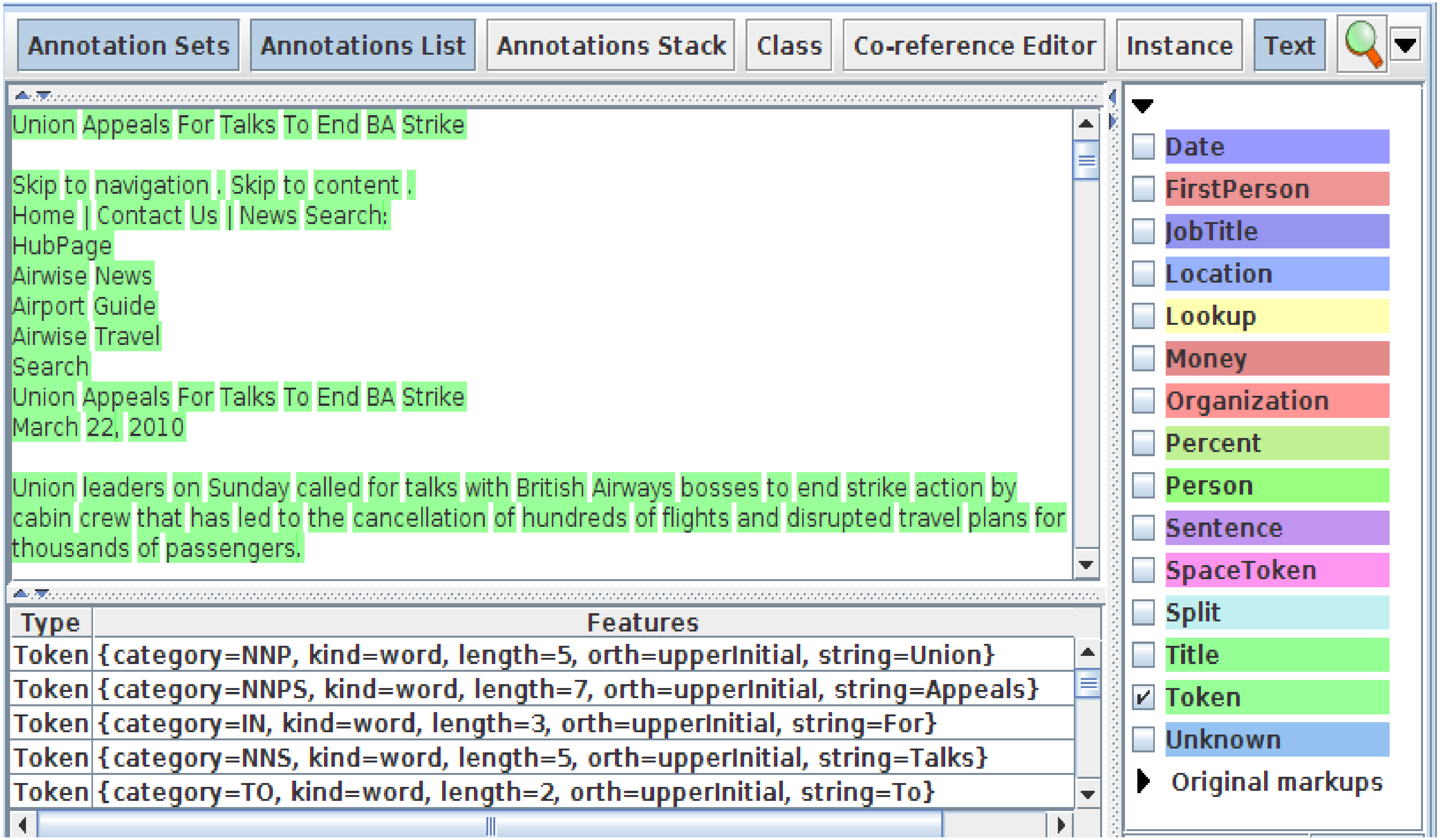
Tokenisation and Sentence Splitting

Tokeniser

- Tokenisation based on Unicode character classes
- Uses 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 wraps the Unicode Tokeniser PR and 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

Re-creating ANNIE

- Tidy up GATE by removing all resources and applications
- Load the news text hands-on corpus
- Create a new application (Conditional 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?

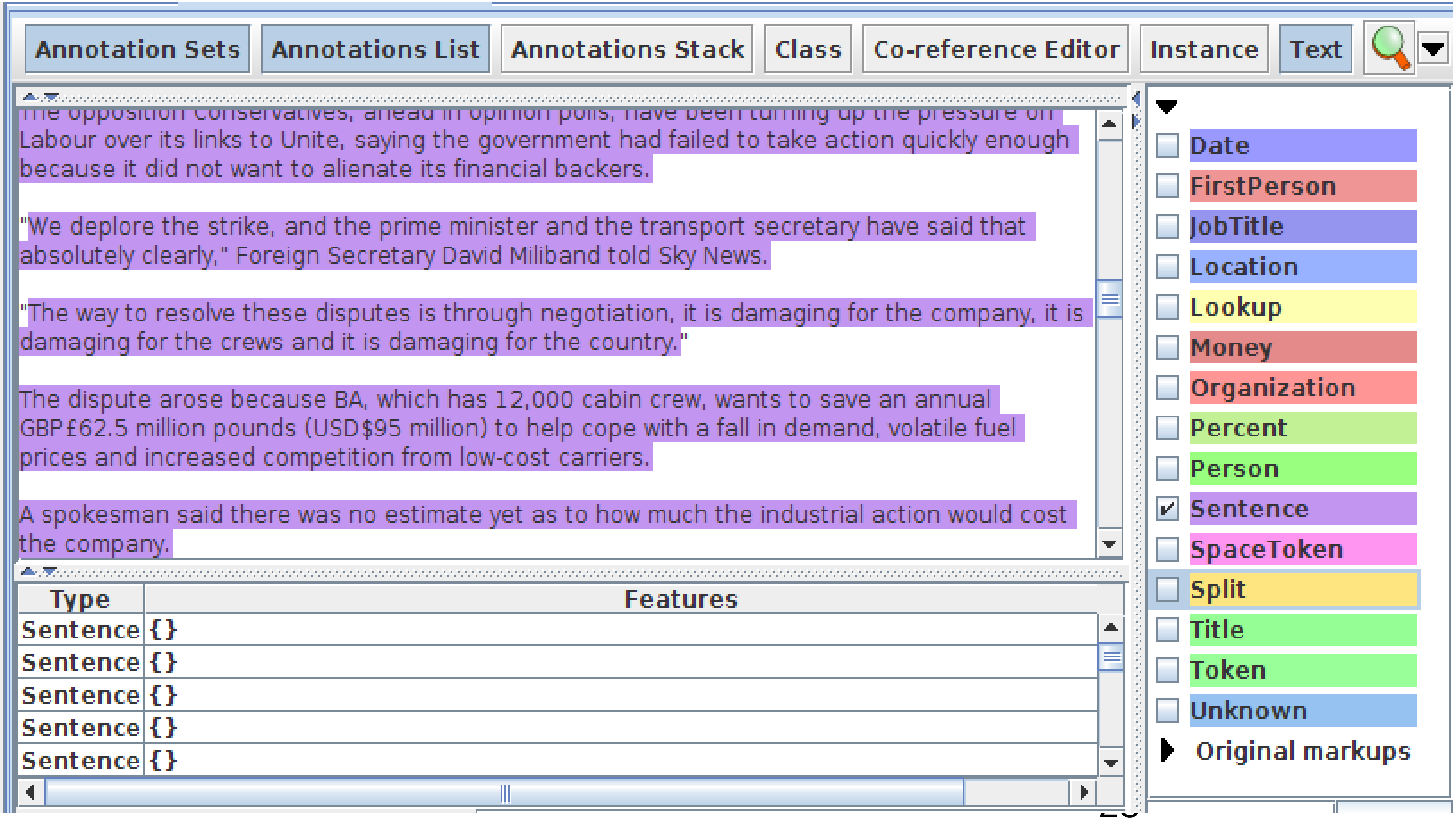
Sentence Splitter

- 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



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

- By default a new-line character always ends a sentence (better suited when importing HTML or PDF)
- Sometimes plain text files have sentences spanning new line characters. In this case it is better to only end a sentence when two consecutive new-line characters are found (paragraph end).
- To do this, create the sentence splitter using “main.jape” instead of “main-single-nl.jape” as the value of the grammar parameter
- 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

Shallow lexico-syntactic features

POS tagger

- Adds category feature to Token annotations
  NNP, VBP, DT, ..
  → Penn Treebank tagset
  https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- ANNIE POS tagger is a Java implementation of Brill's transformation based tagger
- Previously known as **Hepple Tagger** (“heptag”)
- Trained on WSJ
- Default ruleset and lexicon can be modified manually (with a little deciphering)
- 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
- Rule-based: can be modified by the user (instructions in the User Guide)
- Generates “root” feature (lemma) on Token annotations

  “met” → “meet”, “pages” → “page”
- Requires Tokeniser to be run first
- Requires POS tagger to be run first if the considerPOSTag parameter is set to true

Find POS tags and roots

- **Add an ANNIE POS Tagger to your application**
- **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 “category” and “root” have been added

Finding Known Names: Gazetteers

Gazetteers

- Gazetteers are plain text files containing lists of names (e.g rivers, cities, people, …)
- Idea is to quickly find all mentions in a text that match a name
- List is usually compiled into a compact representation for fast matching (e.g. Finite State Machines)
- Each gazetteer has a *definition 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
- Gazetteers generate (by default) Lookup annotations with relevant features corresponding to the list matched
- Lookup annotations are used primarily by the NE transducer

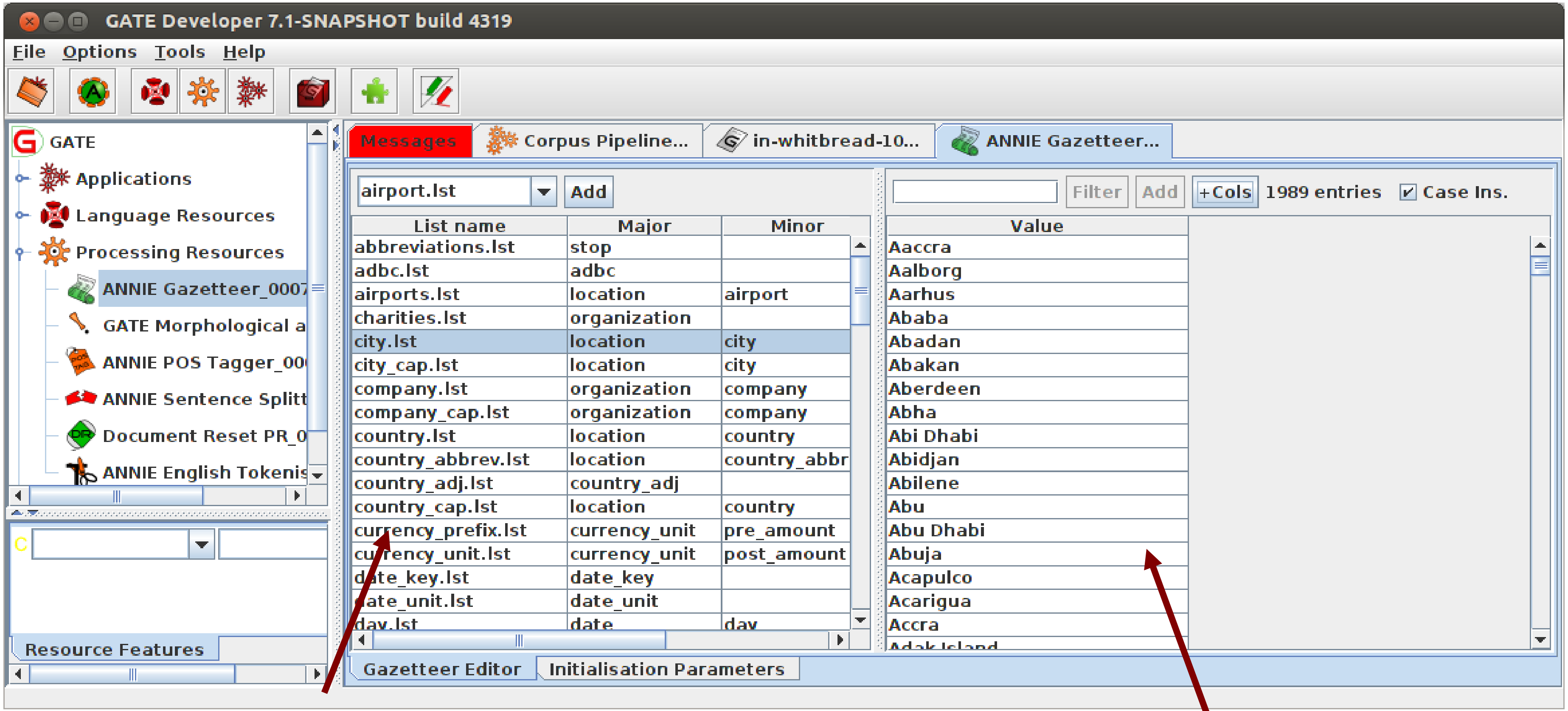
Running the ANNIE Gazetteer

- Various different kinds of gazetteer are available: first we'll look at the default ANNIE gazetteer
- IMPORTANT: there is a copy of the original GATE gazetteer files available in `annie-hands-on/gazetteer` – we use that to avoid modifying the originals!
- **Add the ANNIE Gazetteer PR to the end of your pipeline**
  !! Change the listsURL parameter to the file annie-hands-on/gazetteer/lists.def
- **Re-run the pipeline**
- **Look for “Lookup” annotations and examine their features**

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 (“List name”) you see the definition 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
- Add new list: enter list name in the left pane, click “Add”
- Add new entry: enter entry in the right pane, click “Add”

Gazetteer editor



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_abbr
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

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

lists in the definition file

entries for selected list

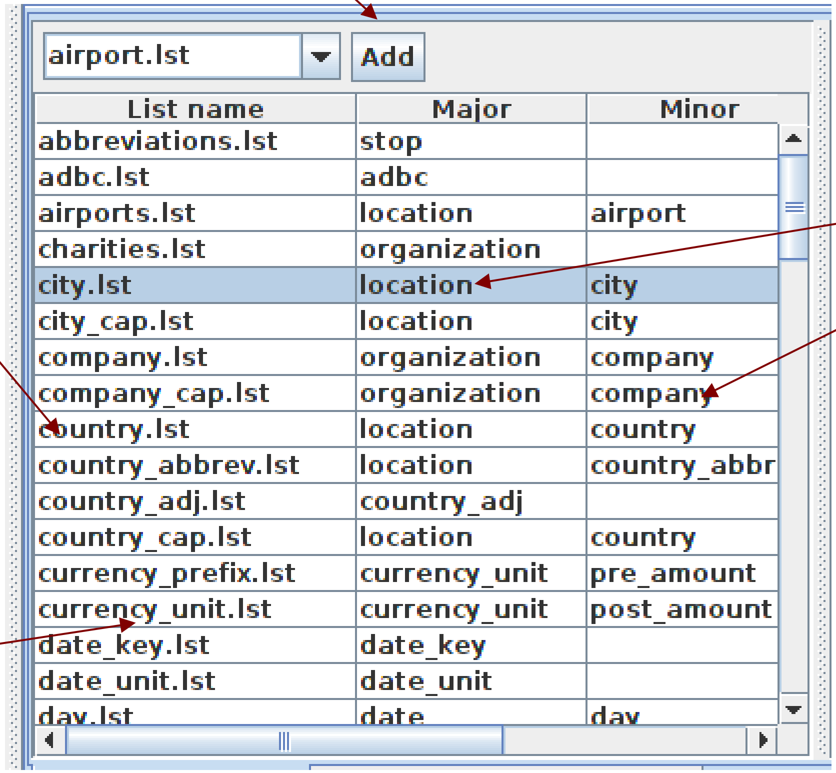
Modifying the definition file

add a new list

edit an existing list name by typing here

edit the major and minor Types by typing here

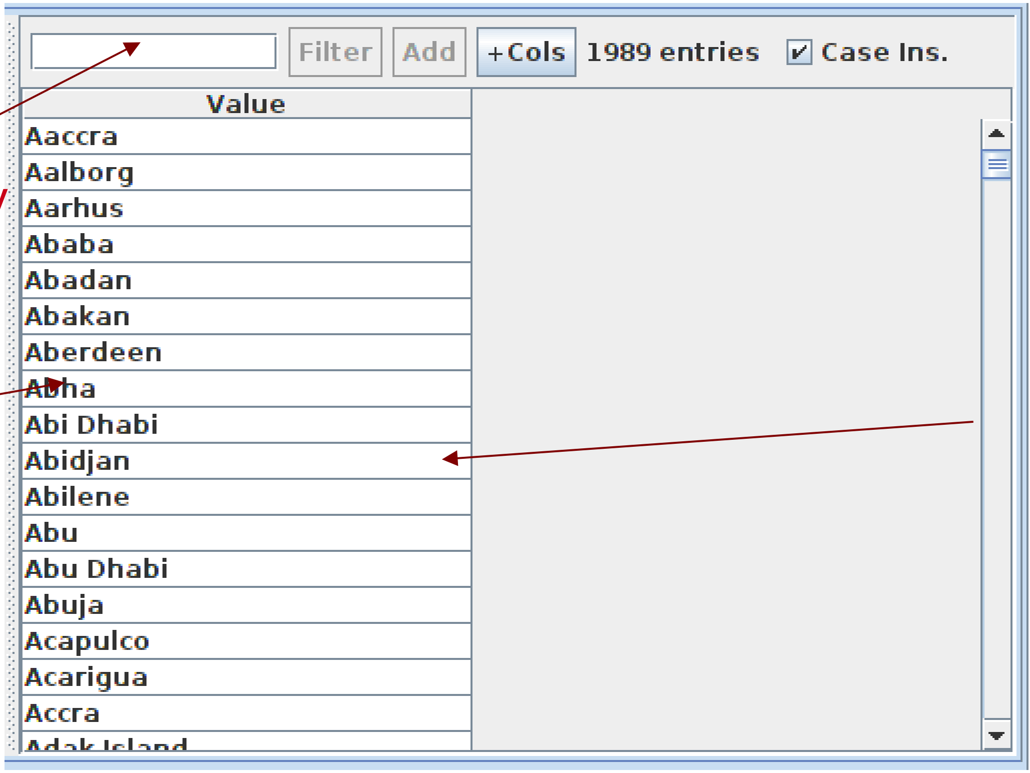
delete a list by right clicking on an entry and selecting Delete



airport.lst ▾ Add

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_abbr
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

Modifying a list



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

add a new entry by typing here

edit an existing entry by typing here

Delete an entry by right clicking and selecting “Delete”

Gazetteer lists

- The ANNIE gazetteer has about 80,000 entries arranged in >100 lists
- Each list is for a category/type, e.g. airports, cities, first names etc.
- List entries might be names or parts of names, 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!

Editing gazetteer lists

- ➢ **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.**
- "Save as" saves the list under the given name plus all lists in the same directory
- ➢ **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”

Per-List Features

- When something in the text matches a gazetteer entry, a Lookup annotation is created, with various features
- The ANNIE gazetteer has the following default feature names: majorType, minorType, language
- These features are used to organize 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"

Per-Entry Features

- Each entry in a list file can have arbitrary features
- For each row must contain the entry to match, then for each feature: a feature separator character, feature name, equals character and feature value, e.g. with the feature separator character “:”

  ```
  Paris:country=France:timezone=CET:lang=fr
  ```

  In many cases, a tab character is the best choice of feature separator character.
- For each match, the per-entry features are added to the Lookup annotation in addition to the per-list features

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”
- Hand-crafted 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 as day name + number + month name
- NE transducer consists of a number of grammars written in the JAPE language
- The next session will be devoted to JAPE

“Transducer”?

- We start with patterns which we want to match in text. The patterns are based on annotations and their features
- Patterns are associated with one or more actions, e.g. create a new annotation for some part of the pattern, add a feature, remove some overlapping annotation
- Patterns and actions are are expressed in the “JAPE” language
  <Pattern> → <Action>
  Rule left hand side: pattern to match
  Rule right hand side: action to carry out
- The JAPE language gets compiled into a FSM for fast matching and carrying out the actions

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 (e.g. 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

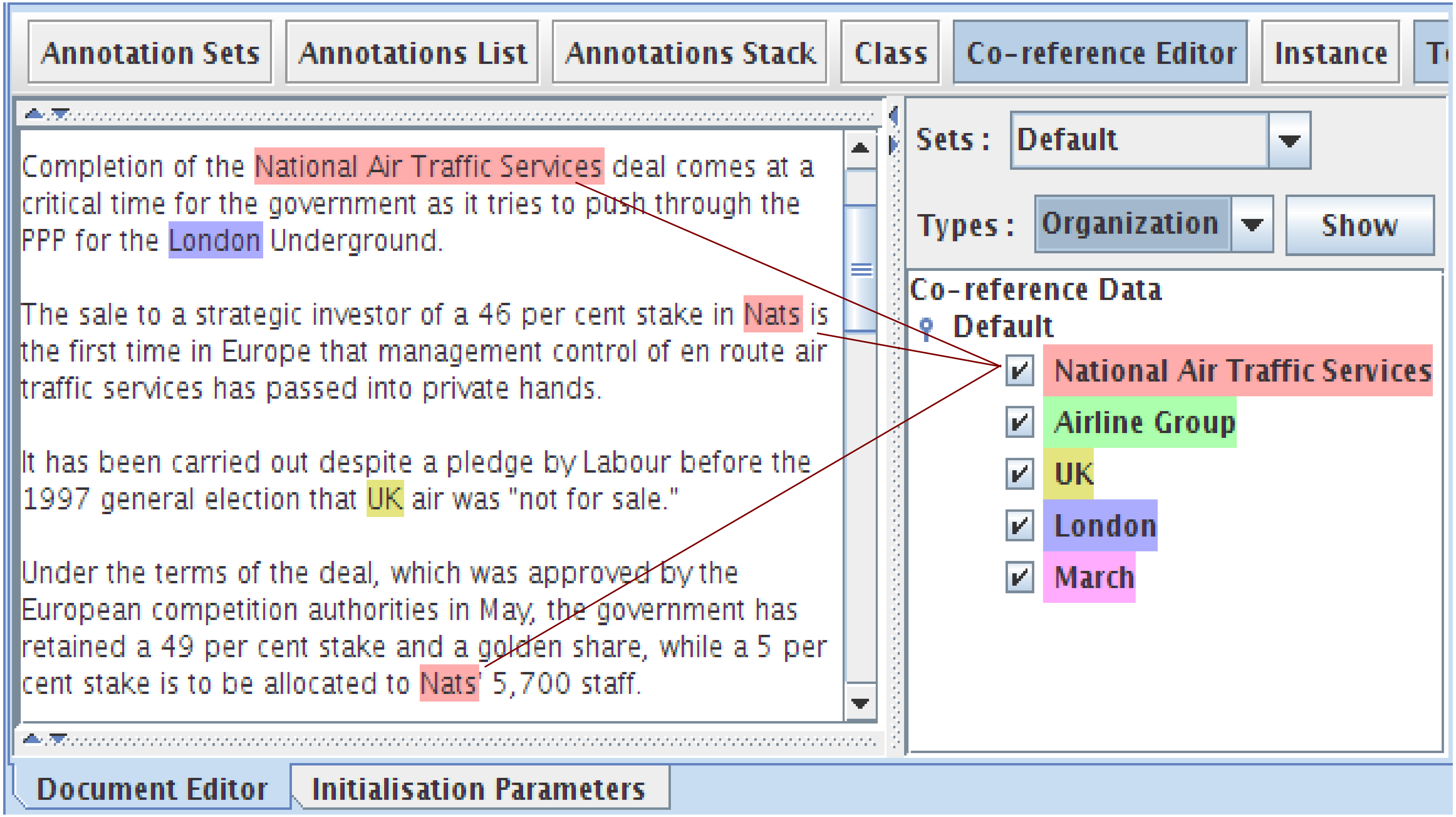
- Performs co-reference resolution based on orthographical information of entities
- Produces a list of annotation IDs that form a co-reference “chain”
- List of such lists stored as a ***document feature*** named “MatchesAnnots”
- Improves results by assigning entity type to previously unclassified 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**
- **Look at the features of NE annotations**
- **Look at the document features in the bottom left pane**
- **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

Co-reference 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

Using the co-reference editor

- **Deselect all items in the co-reference 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 co-references)
- Hovering over a highlighted annotation in the text enables you to add a co-reference between this annotation and one of the co-reference chains listed in the right hand pane
- **Try it!**

ANNIE Alternatives

- ANNIE is a *rule-based* system: manual rules based on linguistic features (e.g. POS tags) and features from gazetteer lists (e.g. minorType=city)
  Advantage: an expert can fine-tune, update, extend incrementally
  Disadvantage: a lot of work, “diminishing results”
- Other approaches uses machine learning
  GATE provides:
  - Stanford NER
  - LingPipe NER
  - OpenNLP NER
  Advantage: no rule crafting necessary
  Disadvantage: need pre-annotated corpus, ML-expert

Evaluation

“We didn’t underperform. You overexpected.”

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 wish we want to evaluate ANNIE. In practice, they could be manual annotations, or annotations from another application.
- **Load the ANNIE system with defaults**
- **Run ANNIE: You should have annotations in the Default set from ANNIE, and in the Key set, against which we can compare them.**

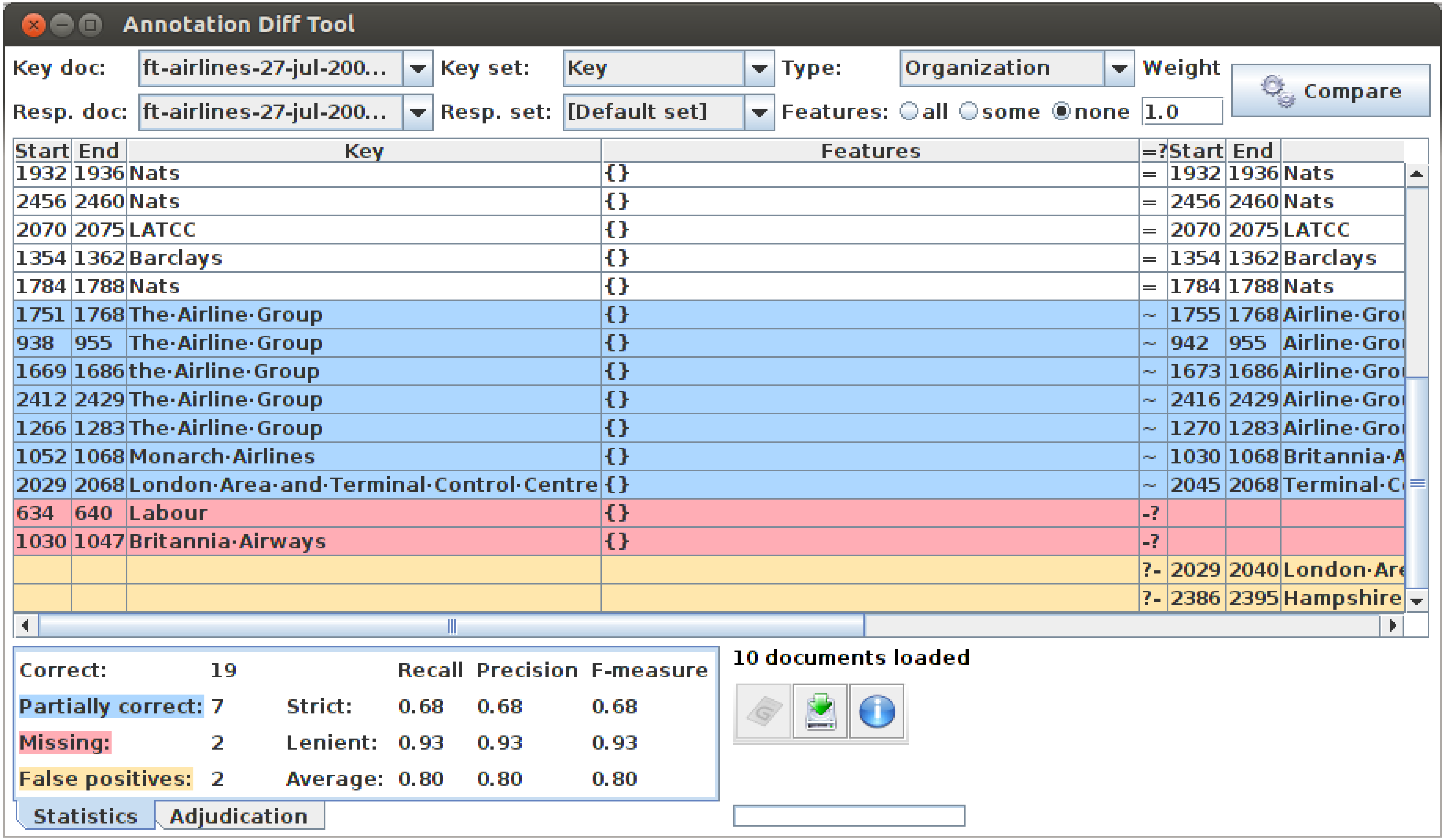
AnnotationDiff

- Graphical comparison of 2 sets of annotations
- Similar to other visual diff tools (kdiff3, tkdiff)
- Compares one document at a time, one annotation type at a time
- Calculates evaluation measures:
  Precision, Recall, F-Measure

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



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

			Recall	Precision	F-measure
Correct:	19				
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

Kinds of Evaluation and 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

Kinds of Evaluation and Terminology

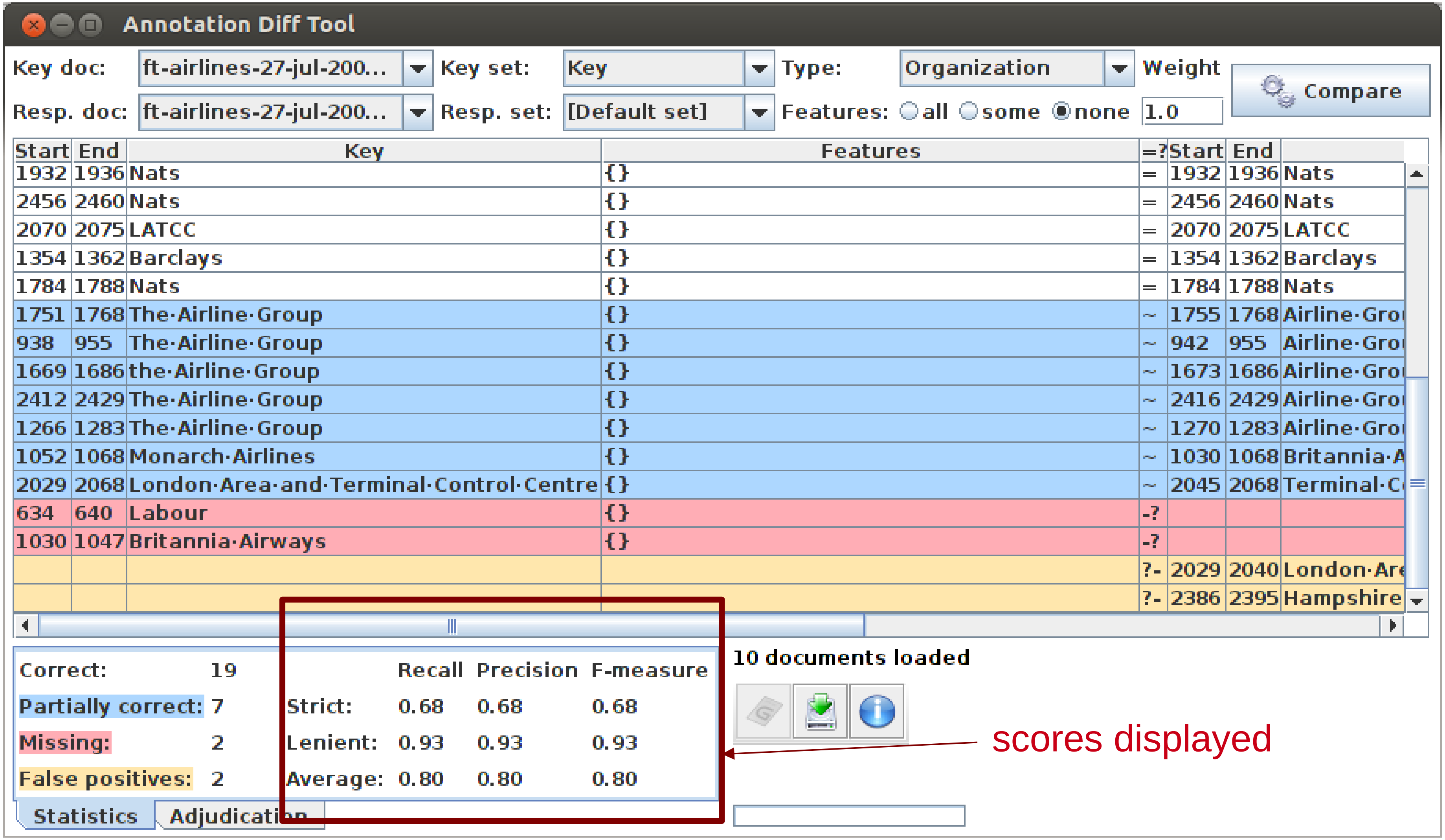
- In NLP we can easily quantify the locations where something should get annotated (Key, Target annotations)
  Harder to quantify the locations where we do not want annotations
  => Tagging
  Measures: Precision, Recall, F-Measure
- [In other situations we have fixed locations and what an algorithm does for that location is either correct or wrong:
  => Classification]
  Measure: Accuracy
- In NLP we sometimes compare the tagging output of two algorithms or two human annotators where no output is "the correct" one.

Tagging Matches

- **(Strict) Correct** = correct type at exact correct position
  (True Positive, TP)
  e.g. annotating “Hamish Cunningham” as a Person
- **Missing** = not annoted (False Negative, FN)
  e.g. not annotating “Sheffield” as a Location
- **Spurious** = wrong type or wrong location (False Positive, FP)
  e.g. annotating “Hamish Cunningham” as a Location

- **Partially correct** = correct type, location overlap
  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

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
Partially correct:	7
Missing:	2
False positives:	2

	Recall	Precision	F-measure
Strict:	0.68	0.68	0.68
Lenient:	0.93	0.93	0.93
Average:	0.80	0.80	0.80

10 documents loaded

scores displayed

NLP Precision

- Fraction of annotations we found that were correct
- Ideally all would be correct, so no spurious!.

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

All the found annotations, the “response” annotations

NLP Recall

- The fraction of the entities that were annotated
- Ideally, all would be correct, so no missing!

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

All the real entities, the “key” annotations or “target” annotations

F-Measure

- Precision and recall tend to trade off against one another
  - Limiting output to only very specific, high confidence annotations will create viewer annotations and thus is likely to reduce recall
  - Creating more generic annotations in an attempt to improve recall is likely to create more spurious ones and reduce precision
- F-measure combines precision and recall into one measure
- Since both precision and recall are fractions the F-measure is the harmonic mean of precision and recall

F-Measure

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

- Sometimes precision or recall is given more weight but usually, precision and recall are equally weighted
- This is also known as F1 or F1.0

F-Measure
Why Harmonic Mean?

$$\mathrm{F}=\frac{1}{\frac{1}{2}\left(\frac{1}{\mathrm{Prec}}+\frac{1}{\mathrm{Rec}}\right)}=\frac{2\,\mathrm{Prec}\cdot\mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}$$

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{Keys}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\rightarrow\frac{1}{\mathrm{Rec}}=\frac{\mathrm{TP}+FN}{TP}=1+\frac{FN}{TP}$$

$$\mathrm{Prec}=\frac{\mathrm{TP}}{\mathrm{Resp}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\rightarrow\frac{1}{\mathrm{Prec}}=\frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}}=1+\frac{\mathrm{FP}}{\mathrm{TP}}$$

$$\rightarrow F=\frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

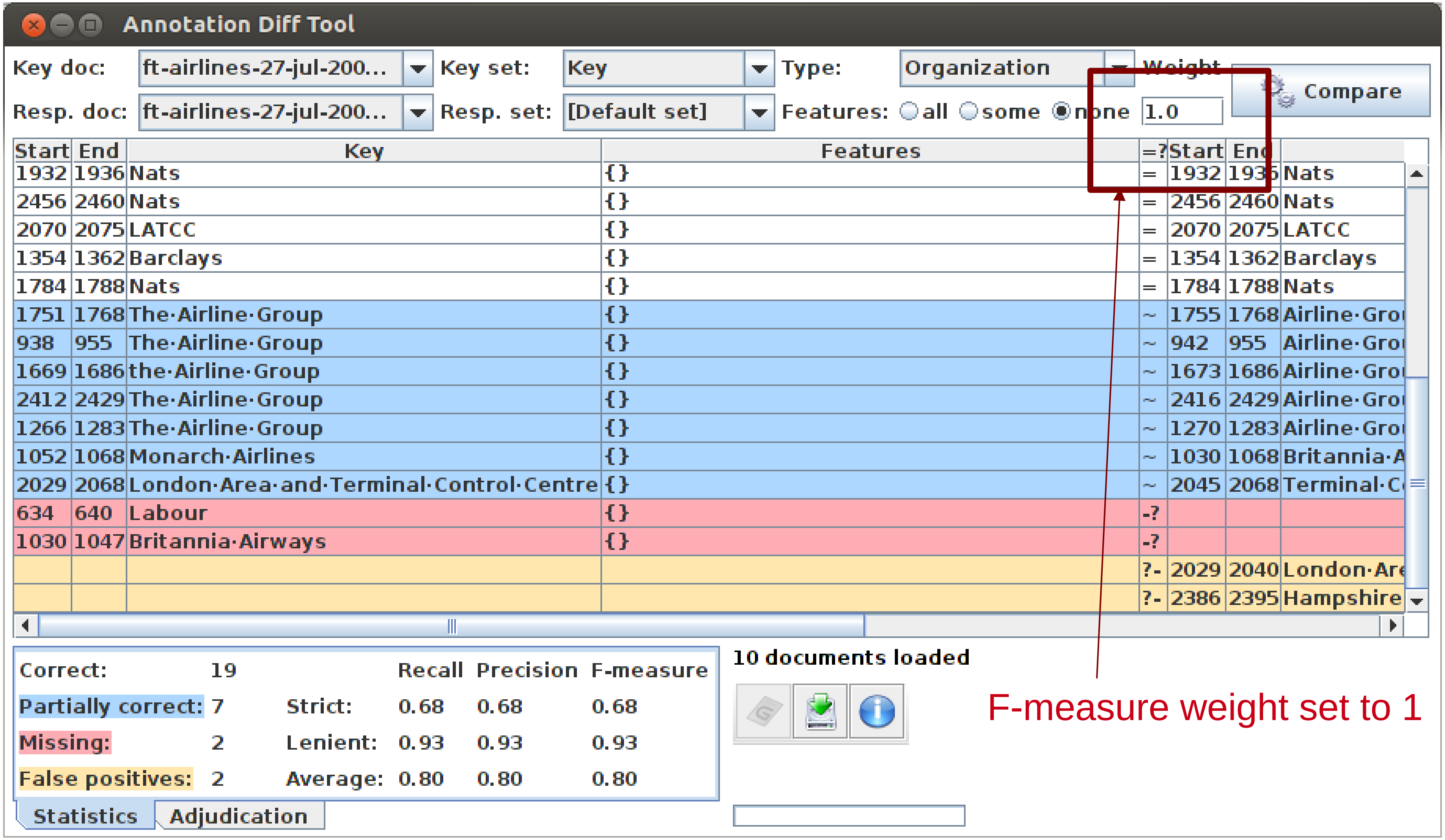
$$F_{\beta}=\frac{(\beta^2+1)\,\mathrm{Prec}\cdot\mathrm{Rec}}{\beta^2\cdot\mathrm{Prec}+\mathrm{Rec}}=\frac{(\beta^2+1)\,\mathrm{TP}}{(\beta^2+1)\,\mathrm{TP}+\mathrm{FP}+\beta^2\,\mathrm{FN}}$$

	Text Mention	Text ¬Mention	
Tagger Mention	TP	FP	Resp onses
Tagger ¬Mention	FN	[TN]	
	Targets, Keys		

Interesting: Errors=FP,FN

TP=True Positive (Correct)
FP=False Positive (Spurious)
FN=False Negative (Missing)
[TN=True Negative (Correct)]

Annotation Diff Defaults to F1



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

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

			Recall	Precision	F-measure
Correct:	19				
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

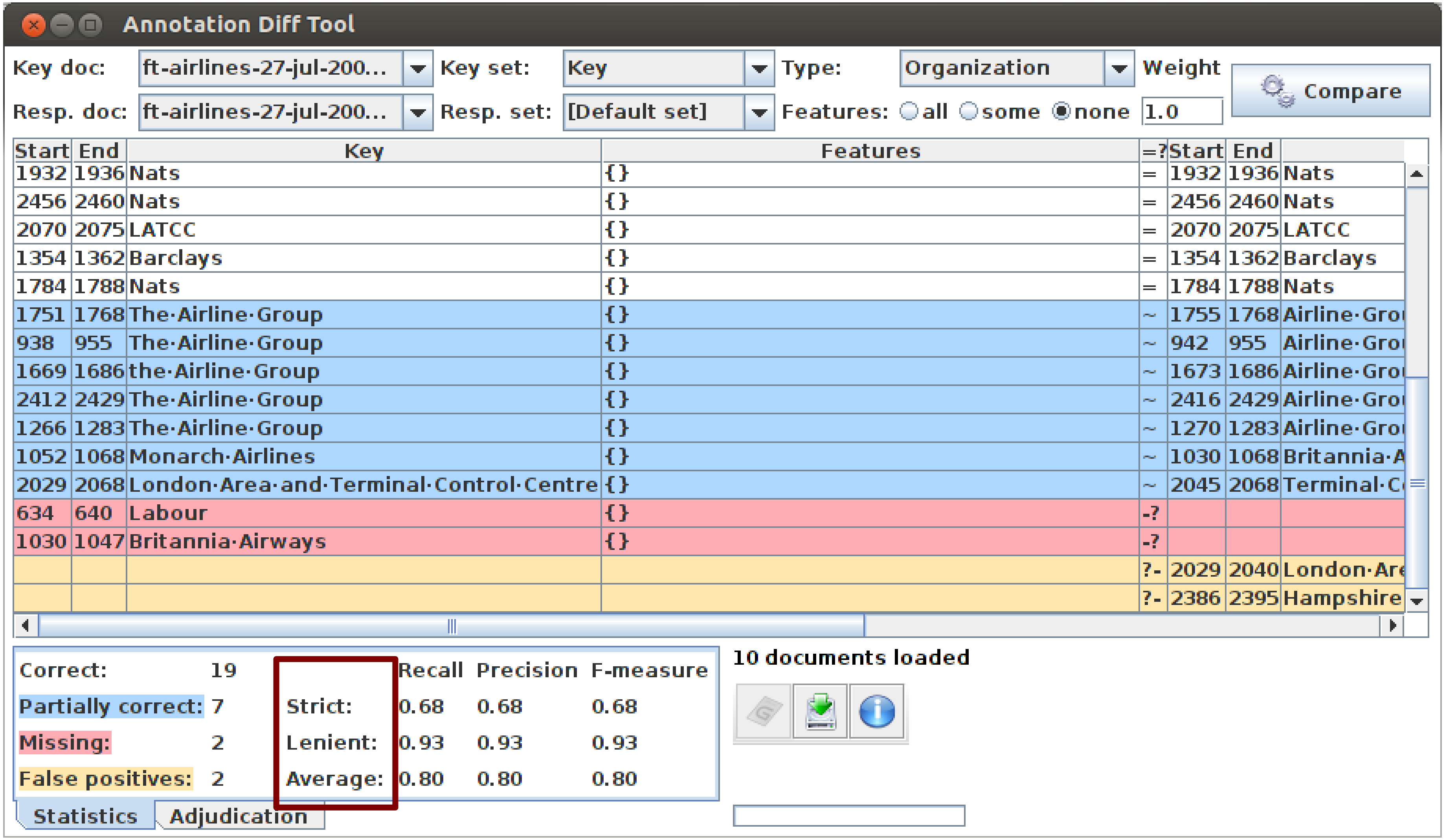
10 documents loaded

F-measure weight set to 1

Count the partial ones?

- How we want to deal with partially correct annotations may differ, depending on our goal
- In GATE, there are 3 different ways to measure them
- Strict measures:
  Only perfectly matching annotations are counted as correct
- Lenient: Partially matching annotations are counted as correct. This usually makes the measures look better
- Average: strict and lenient measures are averaged
  (this is the roughly the same as counting a half weight for every partially correct annotation)

Strict, Lenient, Average



Annotation Diff Tool

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

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·Are
				?-	2386	2395	Hampshire

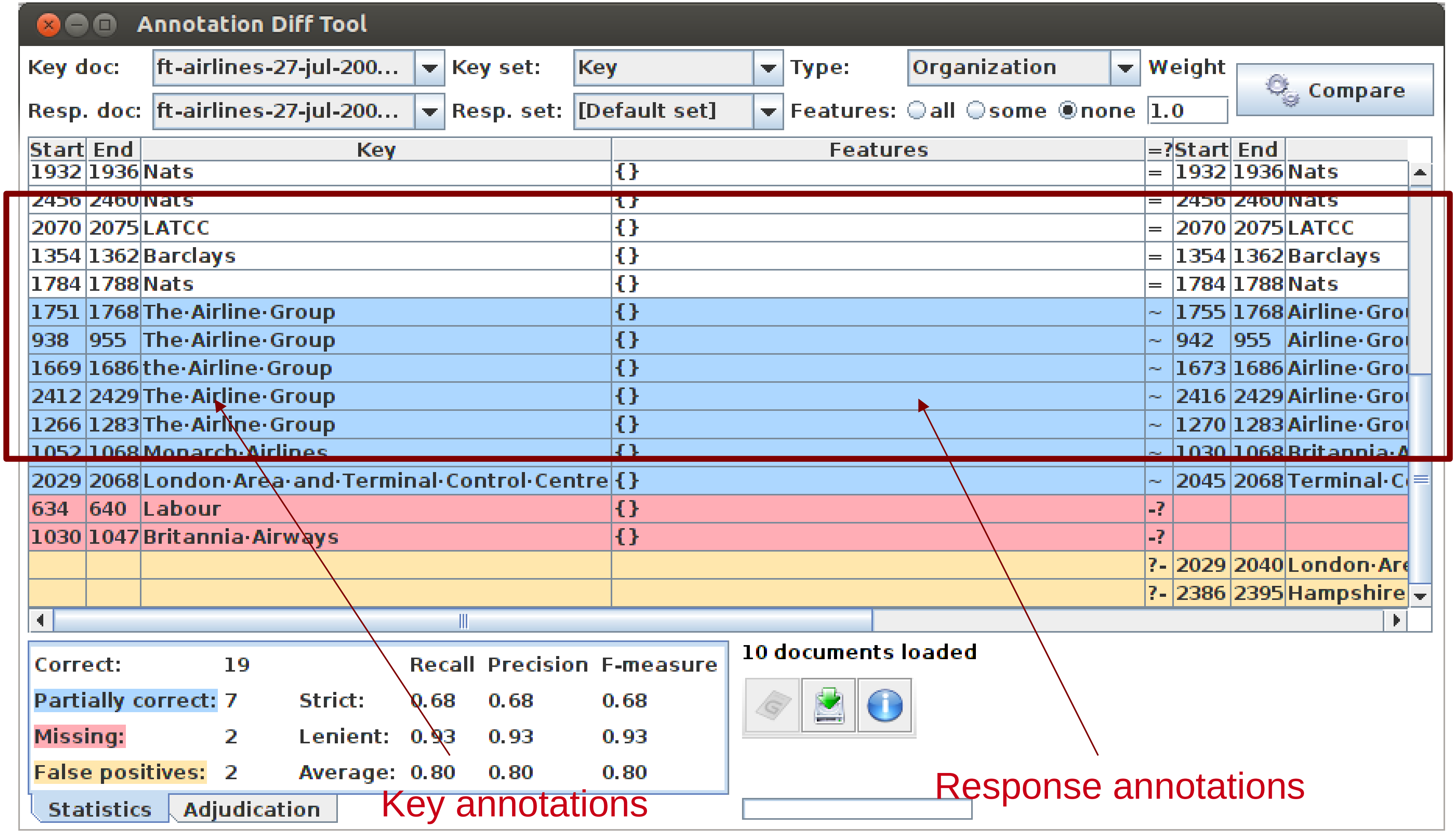
			Recall	Precision	F-measure
Correct:	19				
Partially correct:	7	Strict:	0.68	0.68	0.68
Missing:	2	Lenient:	0.93	0.93	0.93
False positives:	2	Average:	0.80	0.80	0.80

10 documents loaded

Comparing Individual Annotations

- In the AnnotationDiff window, 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



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

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

			Recall	Precision	F-measure
Correct:	19				
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

Key annotations

Response annotations

Corpus Quality Assurance

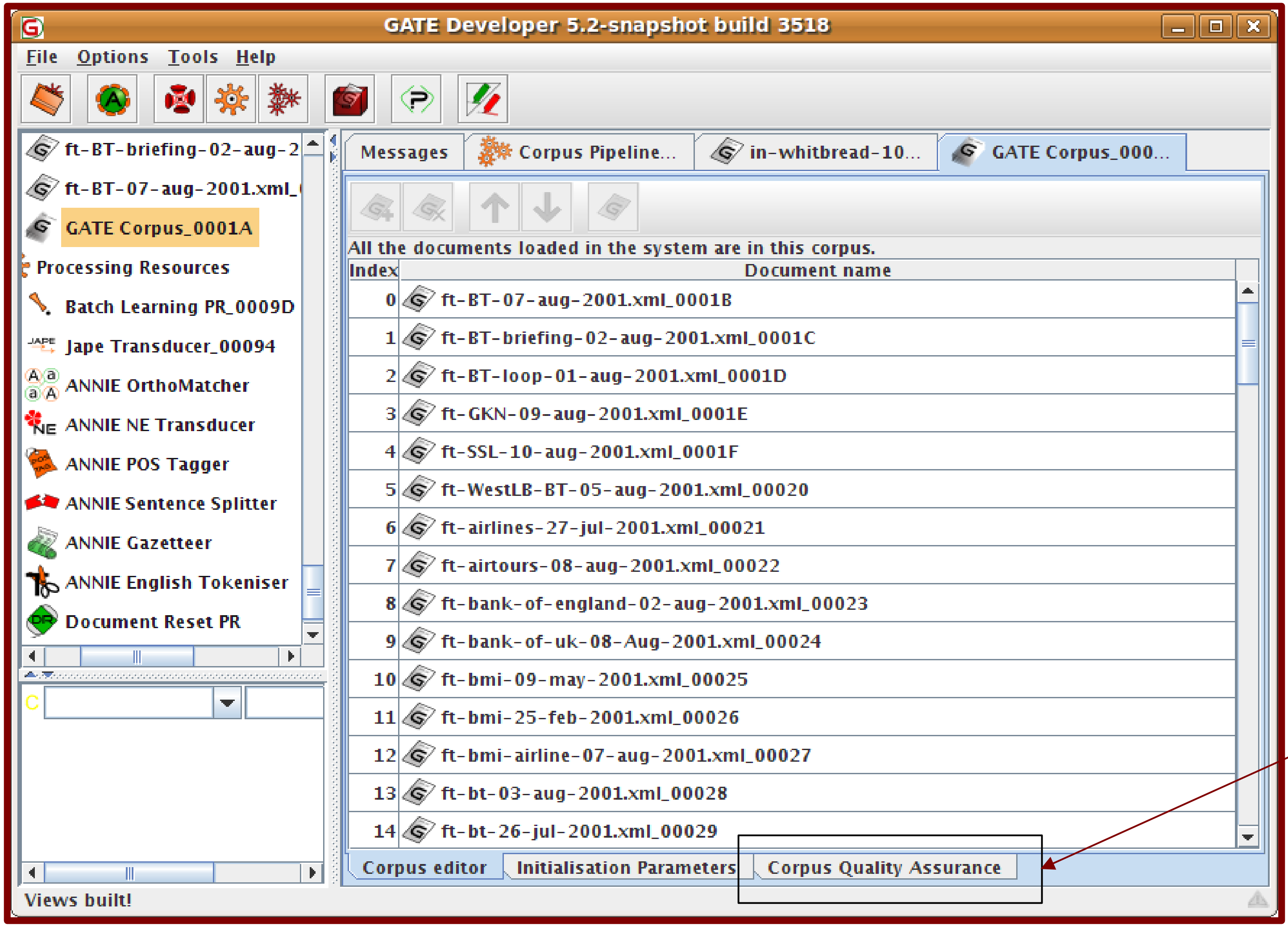
- The Corpus Quality Assurance tool extends the Annotation Diff functionality to an entire corpus
- It produces statistics both for the corpus as a whole (Corpus statistics tab) and for each document separately (Document statistics tab)
- It evaluates several types at once (e.g. Person, Location, Organization)
- It creates **micro averages** and **macro averages** over documents and over types (see next slide)

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 over each of precision, recall and F over all documents or

  over all types: generally less useful / not meaningful
- Over documents, micro average is more useful
- Over types:
  - micro gives the overall picture, more frequent types influence the result more
  - [macro gives ~equal weight to each type]

Try Out Corpus Quality Assurance



GATE Developer 5.2-snapshot build 3518

- ft-BT-briefing-02-aug-2
- ft-BT-07-aug-2001.xml_(
- GATE Corpus_0001A
- Processing Resources
- Batch Learning PR_0009D
- Jape Transducer_00094
- ANNIE OrthoMatcher
- ANNIE NE Transducer
- ANNIE POS Tagger
- ANNIE Sentence Splitter
- ANNIE Gazetteer
- ANNIE English Tokeniser
- Document Reset PR

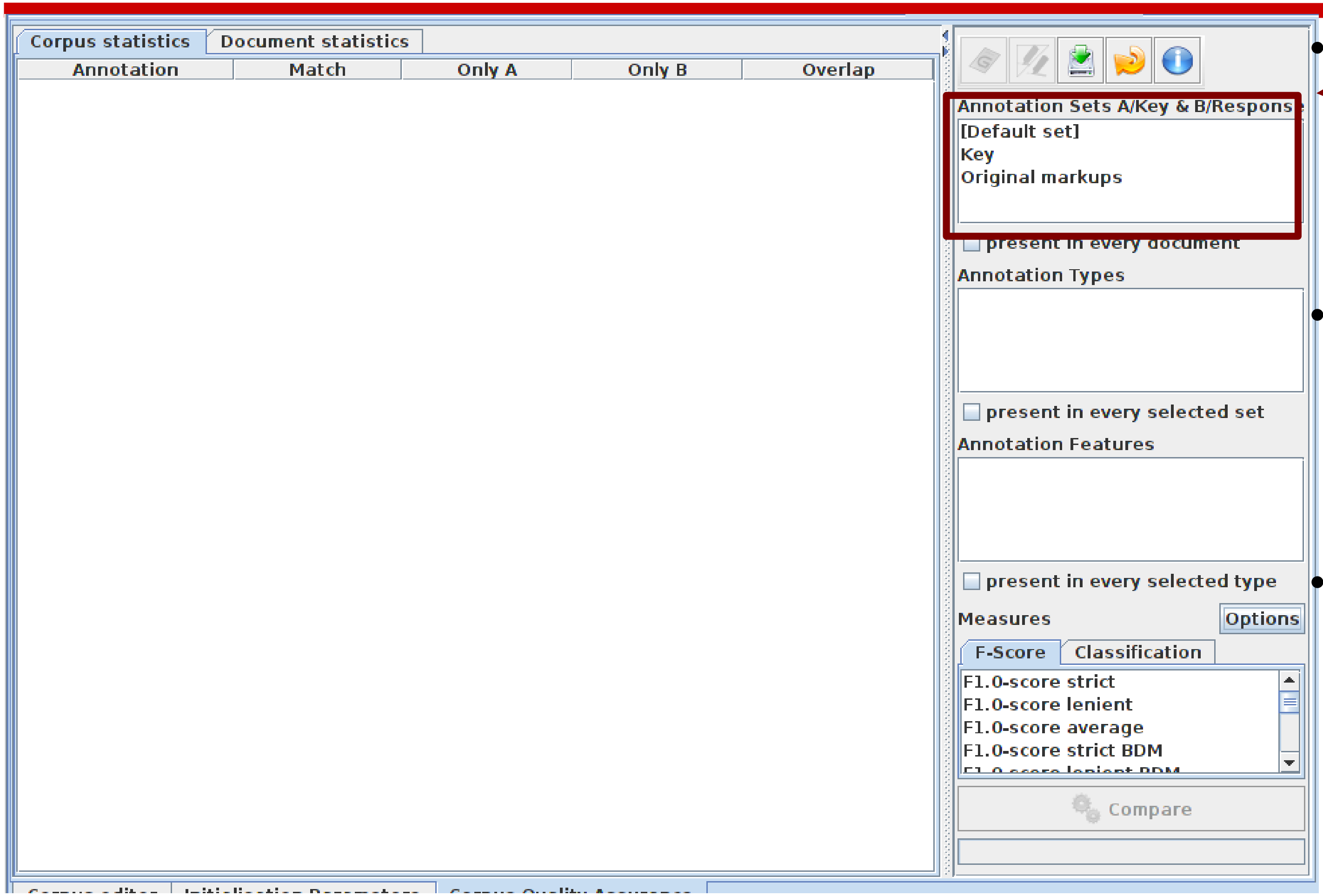
All the documents loaded in the system are in this corpus.

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!

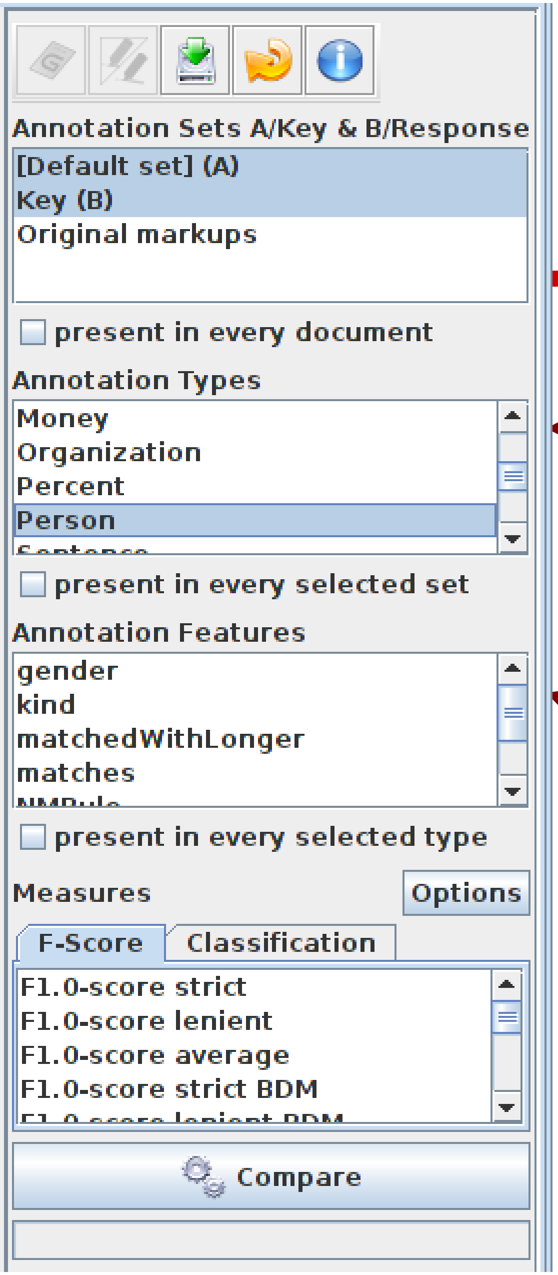
- Open your hands-on corpus and click the Corpus Quality Assurance tab at the bottom of the Display pane.

Select Annotation Sets



- Select the annotation sets you wish to compare.
- Click on 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



Annotation Sets A/Key & B/Response

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

□ present in every document

Annotation Types

Money
Organization
Percent
Person
Sentence

□ present in every selected set

Annotation Features

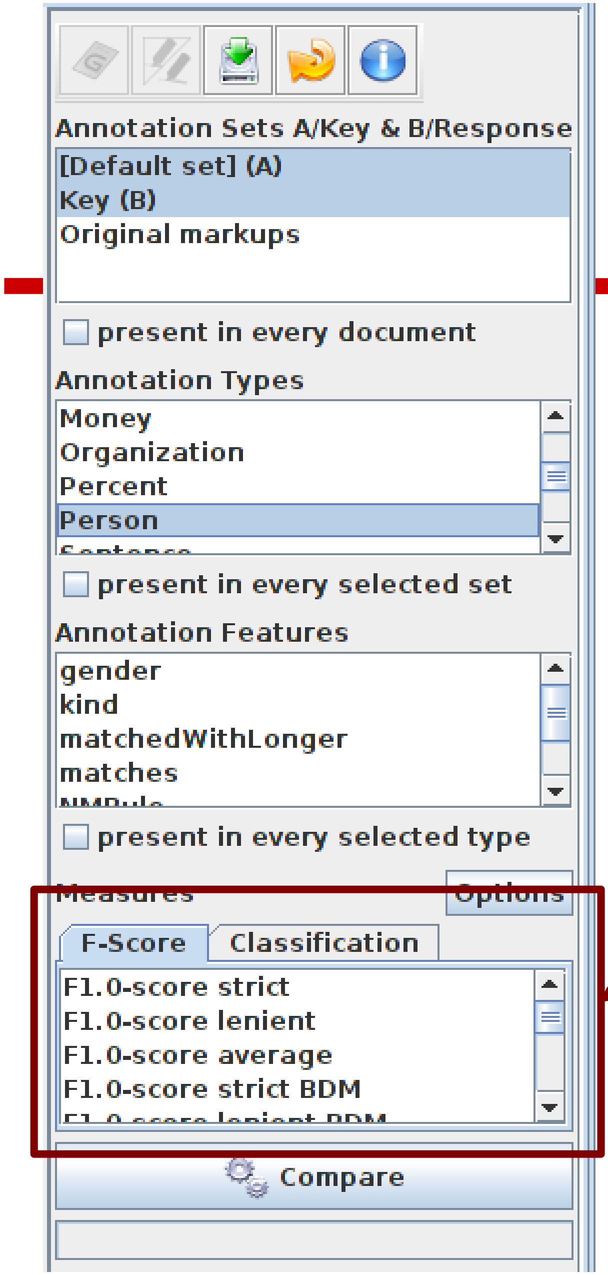
gender
kind
matchedWithLonger
matches
NMRule

□ present in every selected type

- Select the annotation type to compare (suggestion: select Organisation, Person and Location)
- Select the features to include (if any – leave unselected for now)

- You can select as many types and features as you want.

Select measure



Annotation Sets A/Key & B/Response

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

□ present in every document

Annotation Types

Money
Organization
Percent
Person
Sentence

□ present in every selected set

Annotation Features

gender
kind
matchedWithLonger
matches
NMRule

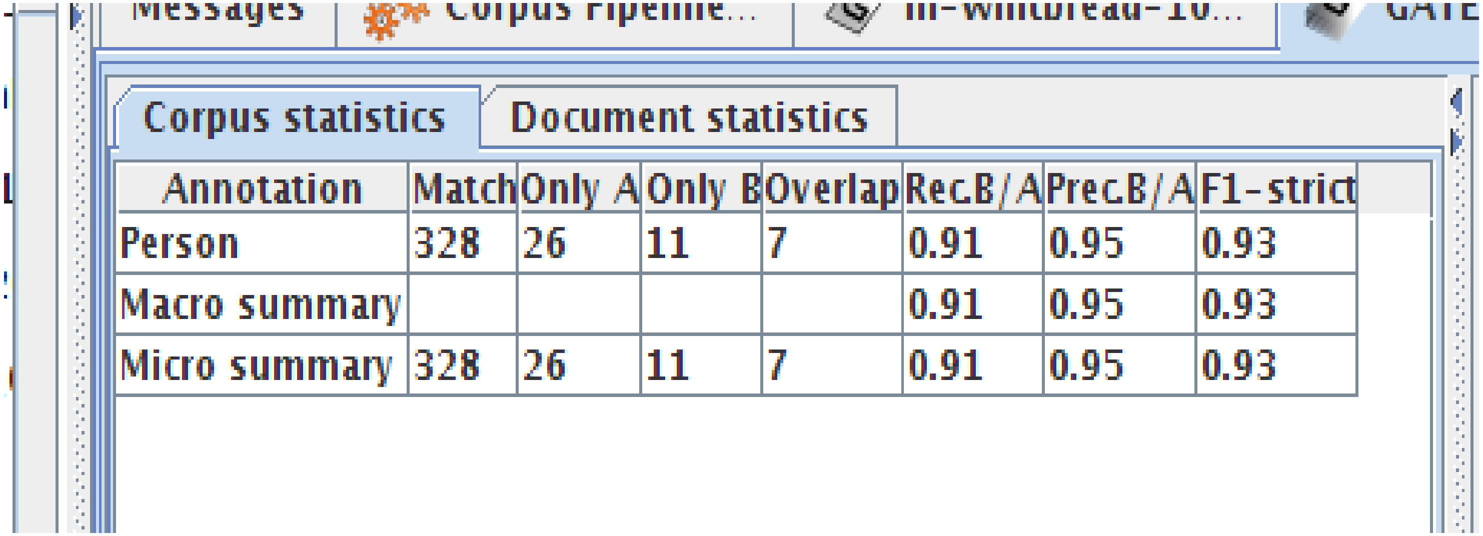
□ present in every selected type

Measures

F1.0-score strict
F1.0-score lenient
F1.0-score average
F1.0-score strict BDM
F1.0-score lenient BDM

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

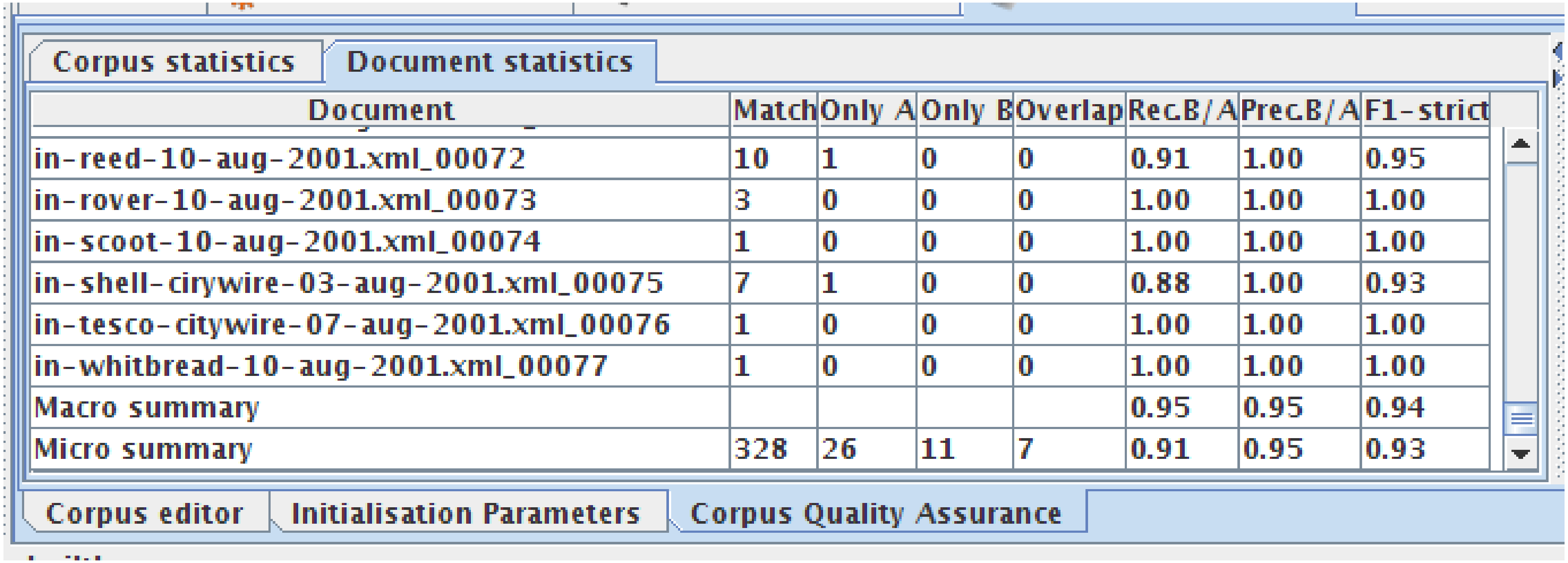
Corpus Statistics Tab



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

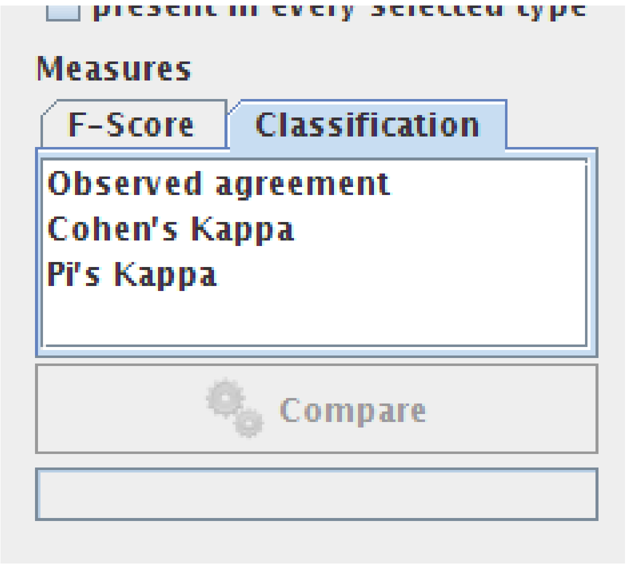
Document Statistics Tab



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

Classification Measures



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

Summary

- This session 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
  - Evaluation using Annotation Diff and Corpus QA

Extra exercises

If you have some spare time, you can try:

- **Load the application** `annie-hands-on/apps/lingpipe-ner.gapp` **Look at the PRs it contains** **Run it on the evaluation corpus, evaluate**
- **Load the application** `annie-hands-on/apps/stanford-ner.gapp` **Look at the PR s it contains** **Run it on the evaluation corpus, evaluate**