BD003 – Introduction to NLP

Part 3: Evaluation

“We didn’t underperform. You overexpected.”

Introduction to Evaluation

- Evaluation of NLP tools is very important because we need to know how well our tools are performing
- Is it actually worth developing an automatic tool to perform a task?
- Especially in GATE, there is often a choice of which tool to use for a job (e.g. multiple parsers) so we might want to know which one is best
- We need to know whether changes we make to the tools will improve or harm our system: e.g. making components case-insensitive might improve Recall but harm Precision
- We will look at what evaluation metrics to use for NLP, and some tools to perform evaluation

Evaluation exercises: preparation

- Restart GATE, or close all documents and PRs to tidy up
- Load the ANNIE 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
- Visual diff representation, like tkdiff
- Compares one document at a time, one annotation type at a time

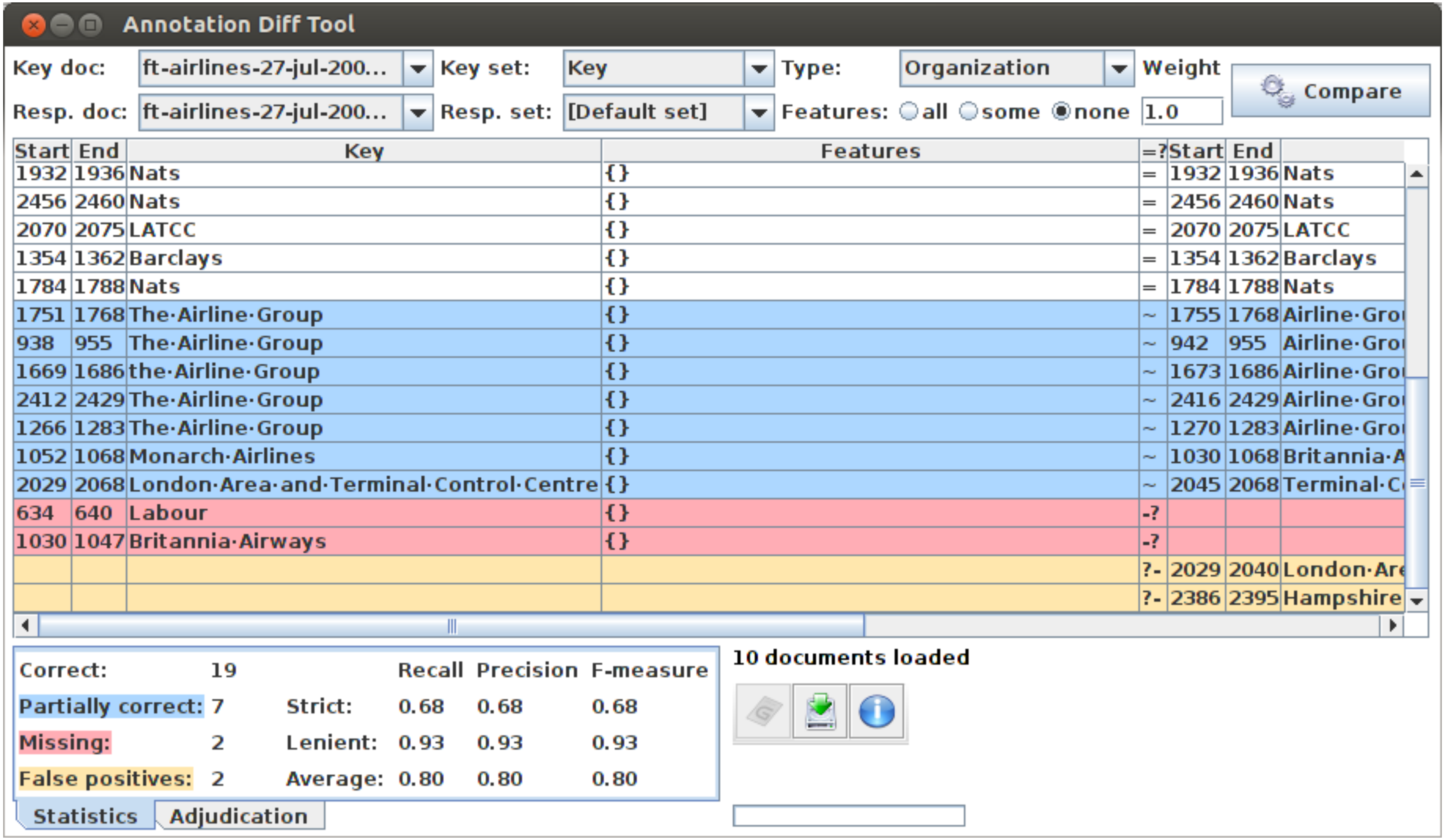
Annotations are like squirrels…

Annotation Diff helps with “spot the difference”

Annotation Diff Exercise

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

Annotation Diff



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

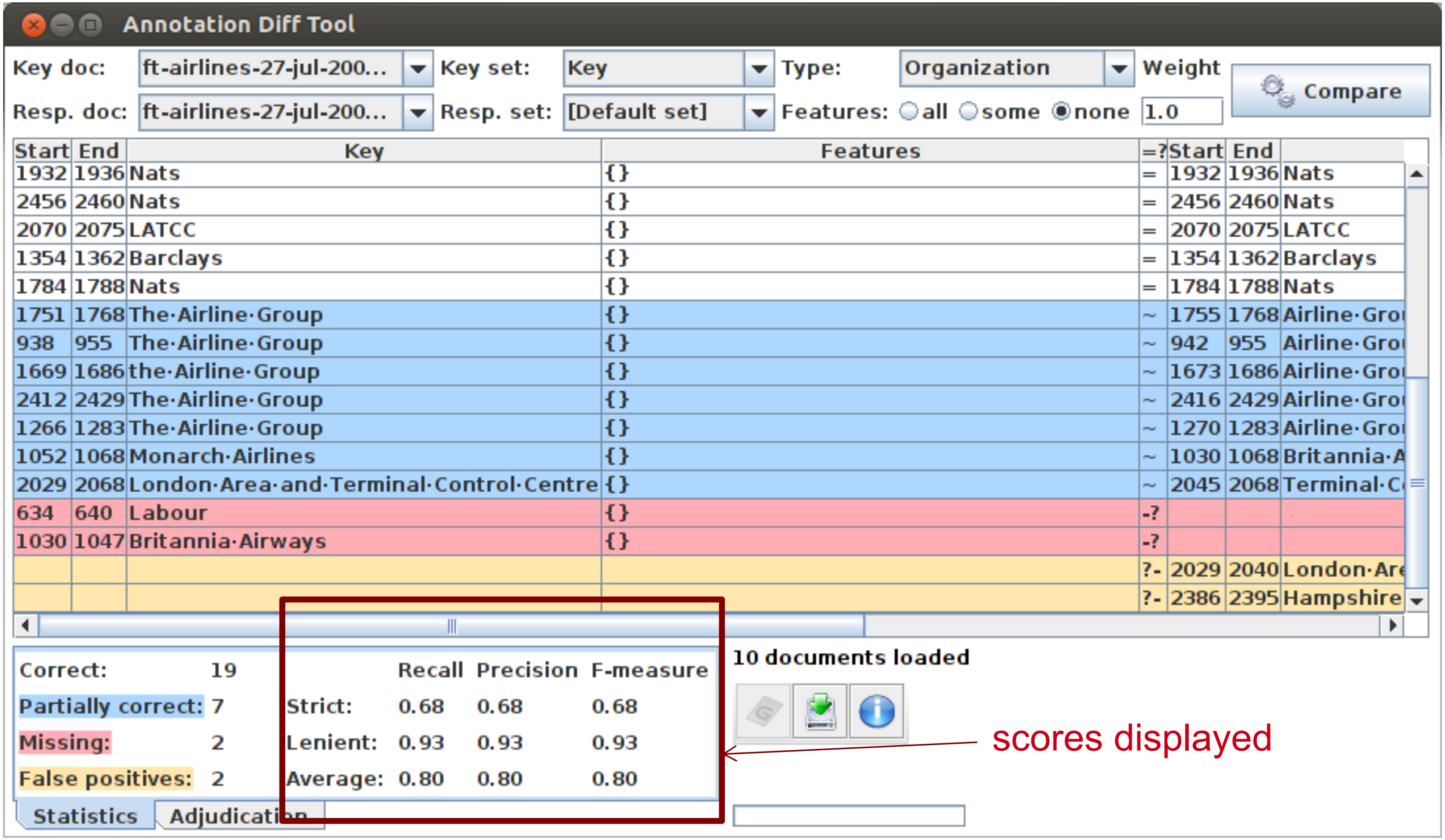
A Word about Terminology

- 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

Measuring success

- In IE, we classify the annotations produced in one of 4 ways:
- **Correct** = things annotated correctly
  - e.g. annotating “Donald Trump” 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 “London” as a Location in “London Traffic Centre”
- **Partially correct** = the annotation type is correct, but the span is wrong
  - e,g, annotating just “Trump” as a Person (too short) or annotating “Unfortunately Donald Trump” 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

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

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

Statistics Adjudication

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

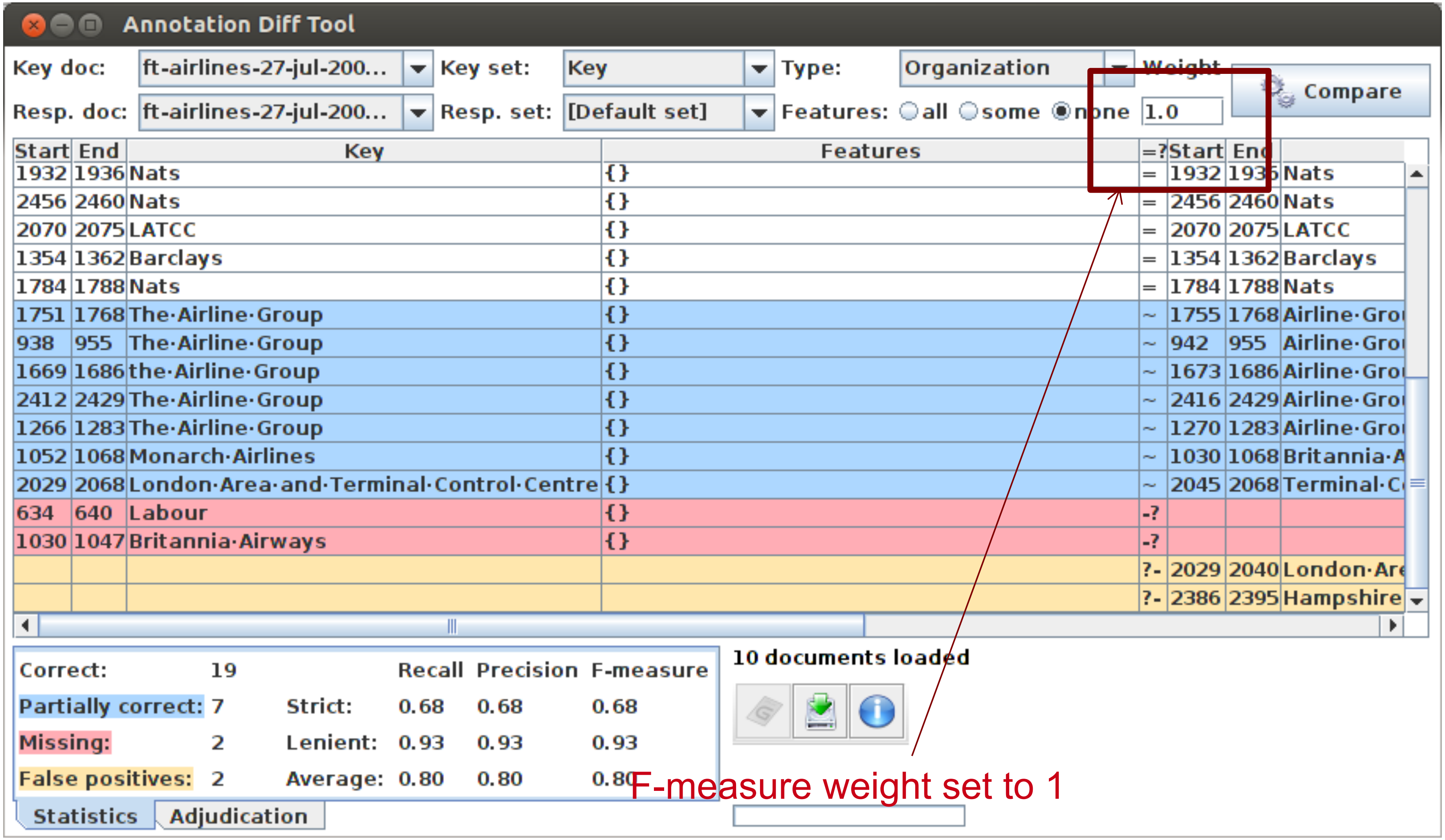
- Precision and recall tend to trade off against one another
- If you specify your rules precisely to improve precision, you may get a lower recall
- If you make your rules very general, you get good recall, but low precision
- This makes it difficult to compare applications, or to check whether a change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



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	0.8

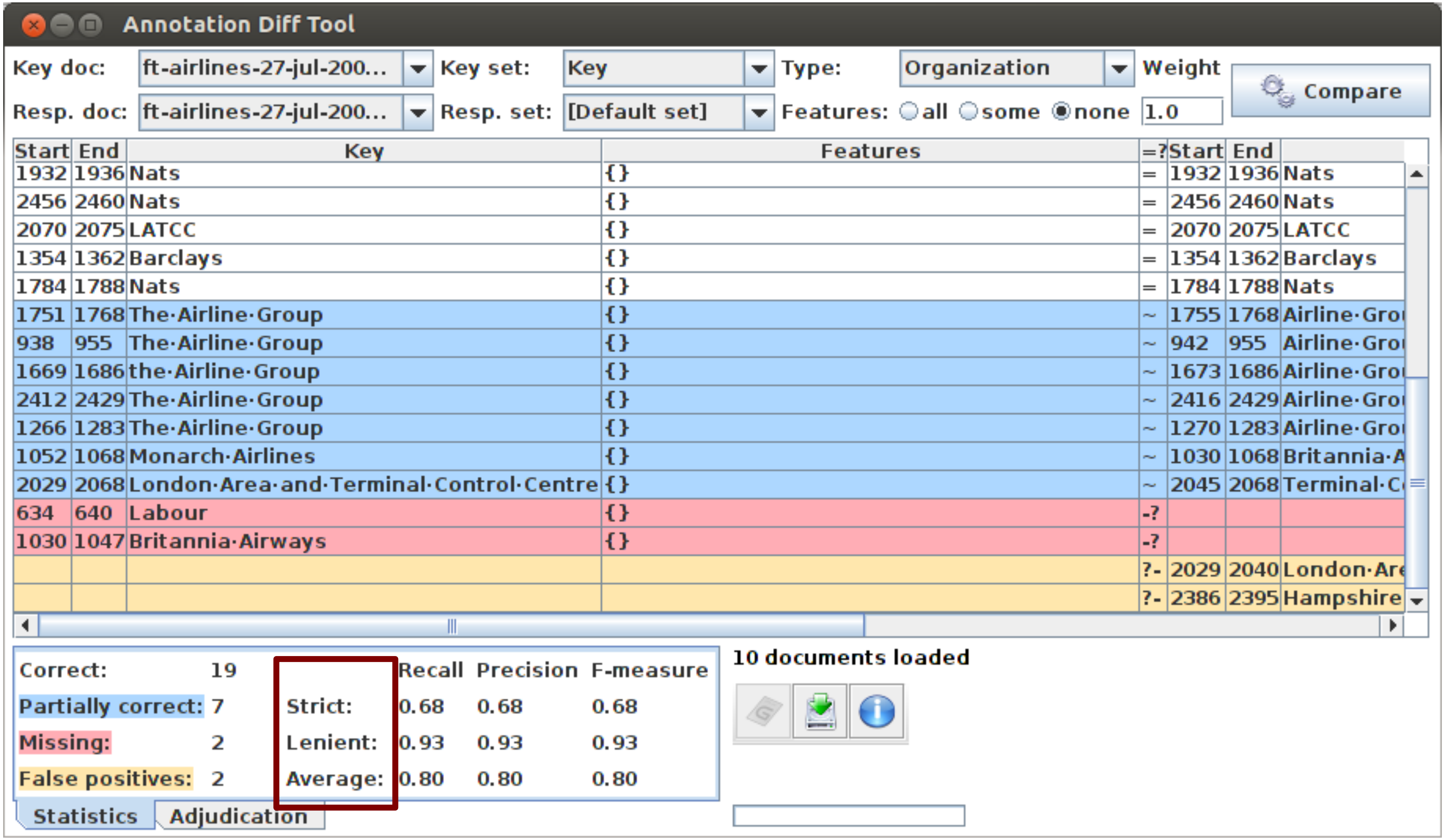
10 documents loaded

F-measure weight set to 1

How to evaluate partially correct annotations

- 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 :-)
- We might use Lenient when the span of the annotation isn’t so important

Strict, Lenient and Average



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

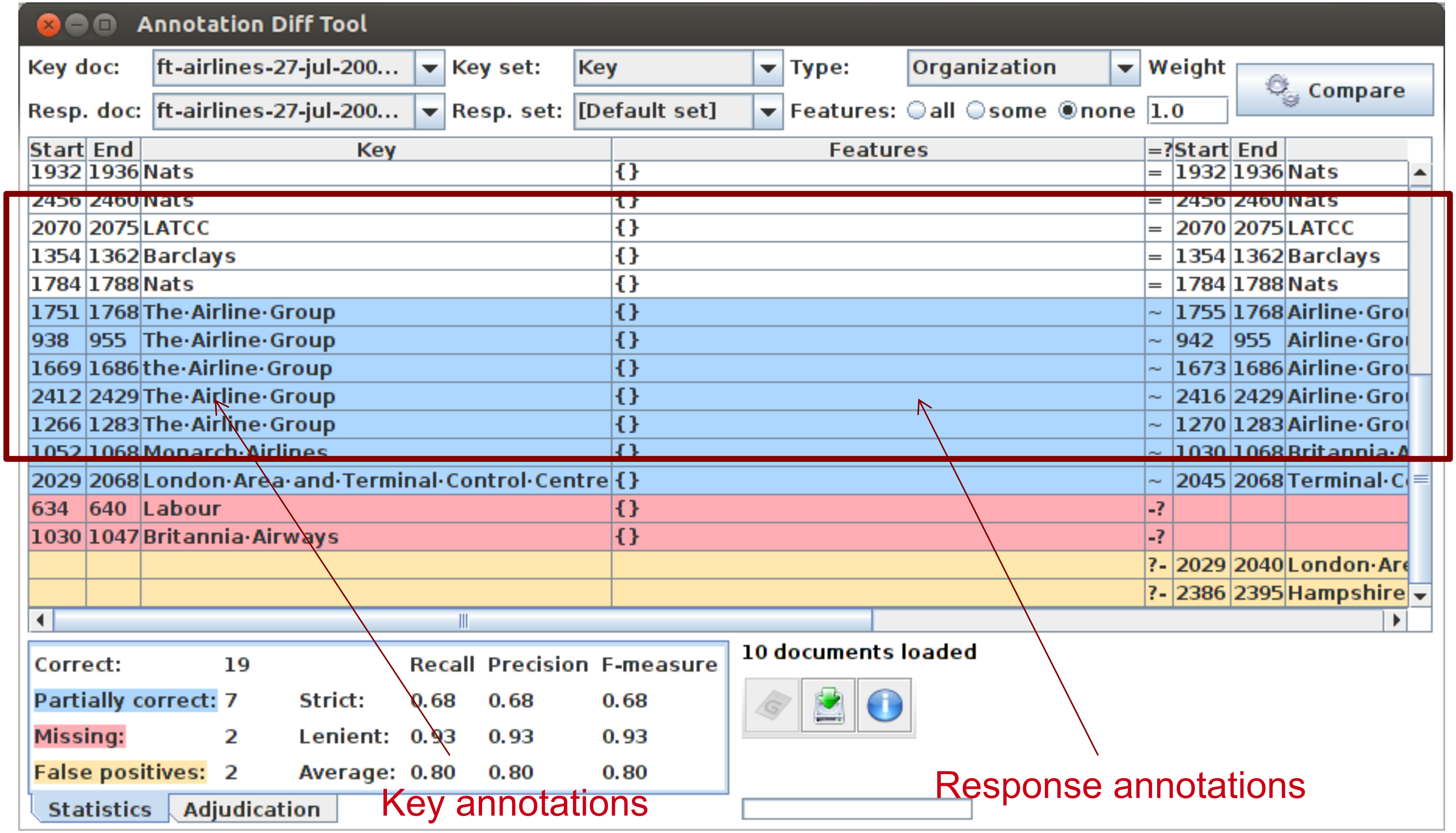
			Recall	Precision	F-measure
Correct:	19				
Partially correct:	7	Strict:	0.68	0.68	0.68
Missing:	2	Lenient:	0.93	0.93	0.93
False positives:	2	Average:	0.80	0.80	0.80

10 documents loaded

Comparing the individual annotations

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

Comparing the annotations



Annotation Diff Tool

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

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

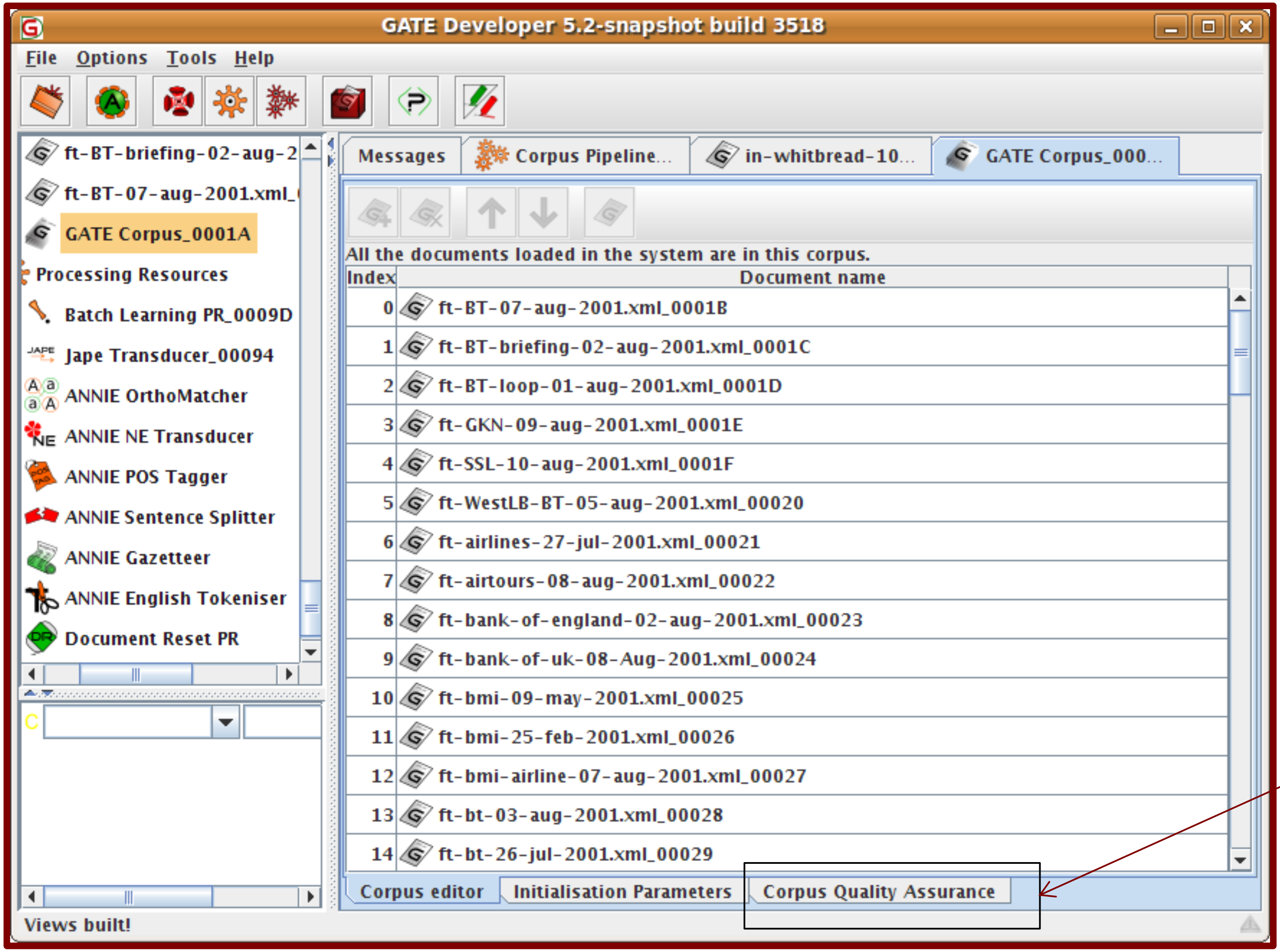
Key annotations

Response annotations

Corpus Quality Assurance

- 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



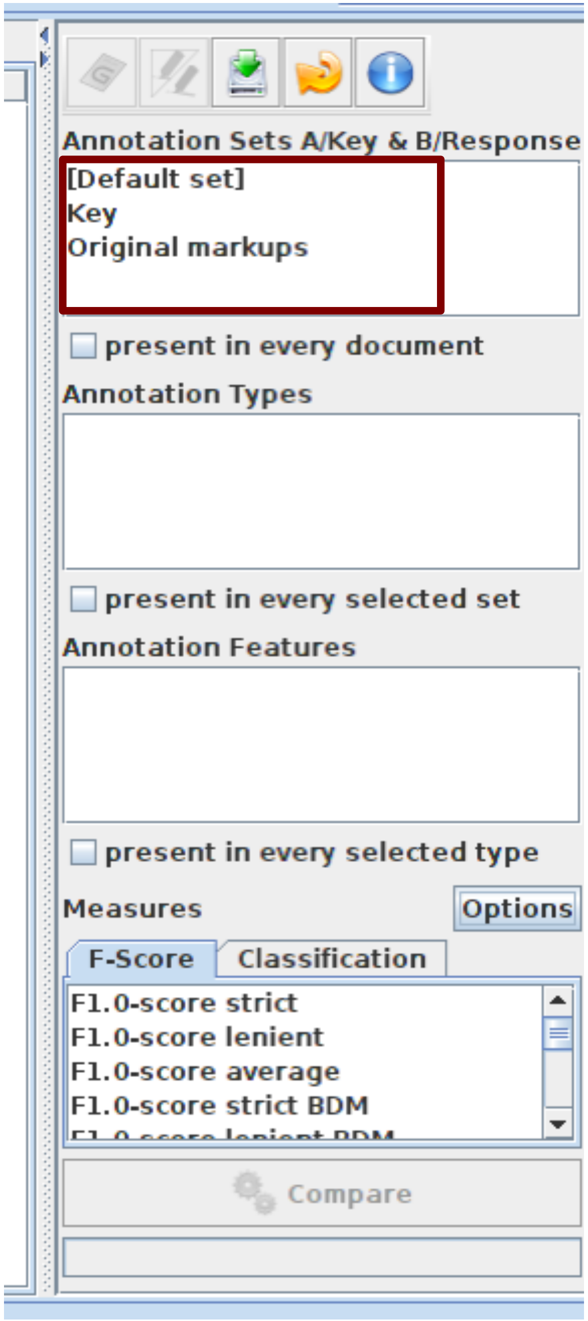
GATE Developer 5.2-snapshot build 3518

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

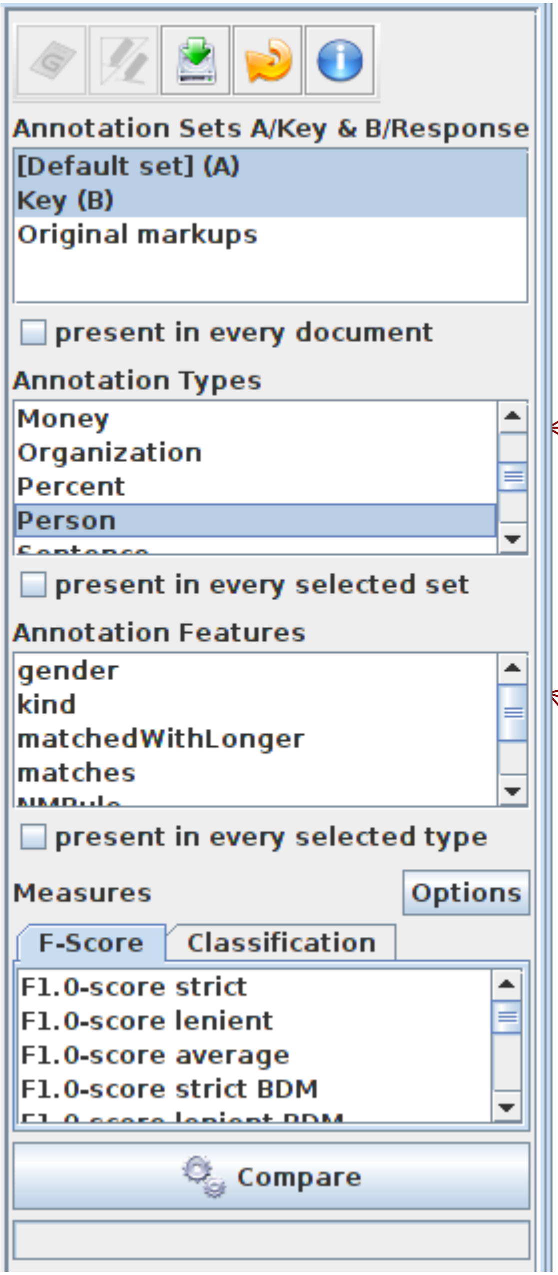
Index	Document name
0	ft-BT-07-aug-2001.xml_0001B
1	ft-BT-briefing-02-aug-2001.xml_0001C
2	ft-BT-loop-01-aug-2001.xml_0001D
3	ft-GKN-09-aug-2001.xml_0001E
4	ft-SSL-10-aug-2001.xml_0001F
5	ft-WestLB-BT-05-aug-2001.xml_00020
6	ft-airlines-27-jul-2001.xml_00021
7	ft-airtours-08-aug-2001.xml_00022
8	ft-bank-of-england-02-aug-2001.xml_00023
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

- Open yo hands-o and clic Corpus Assuran the botto Display

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.


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

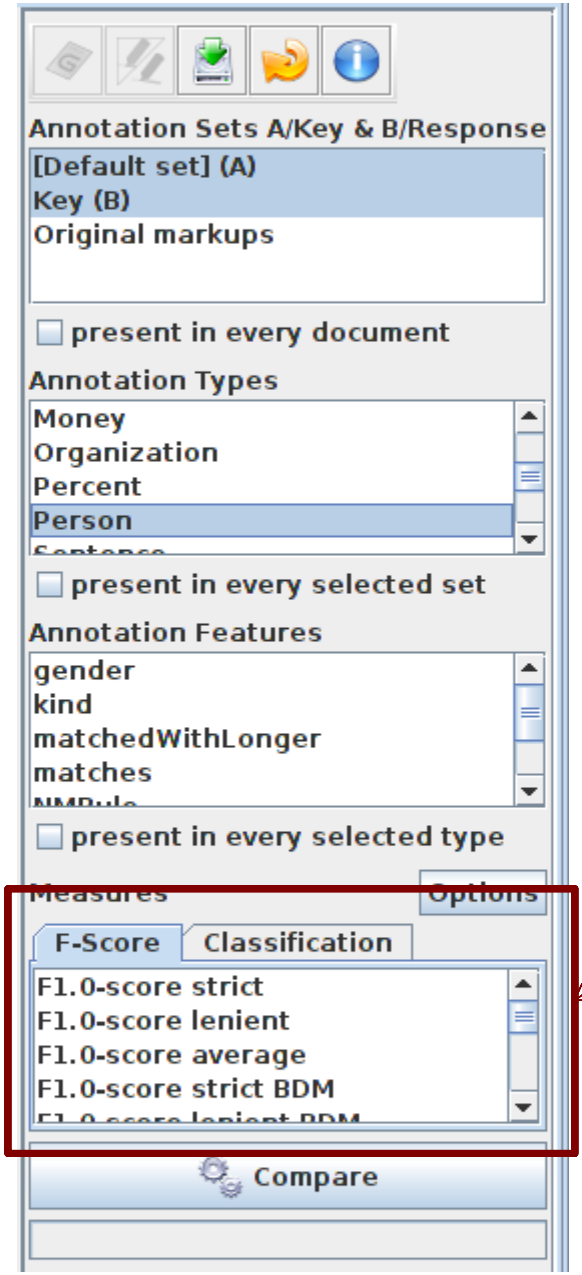
Measures

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

Select Type

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

Select measure



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

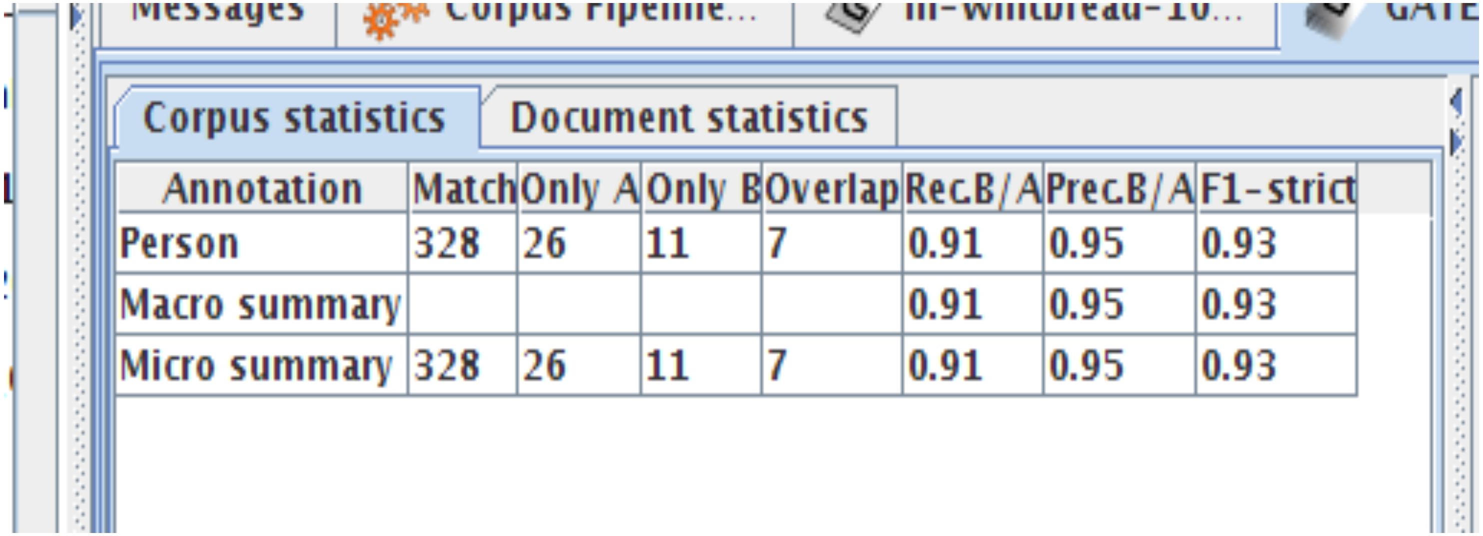
□ present in every selected type

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.
- Select Compare

Corpus Statistics Tab



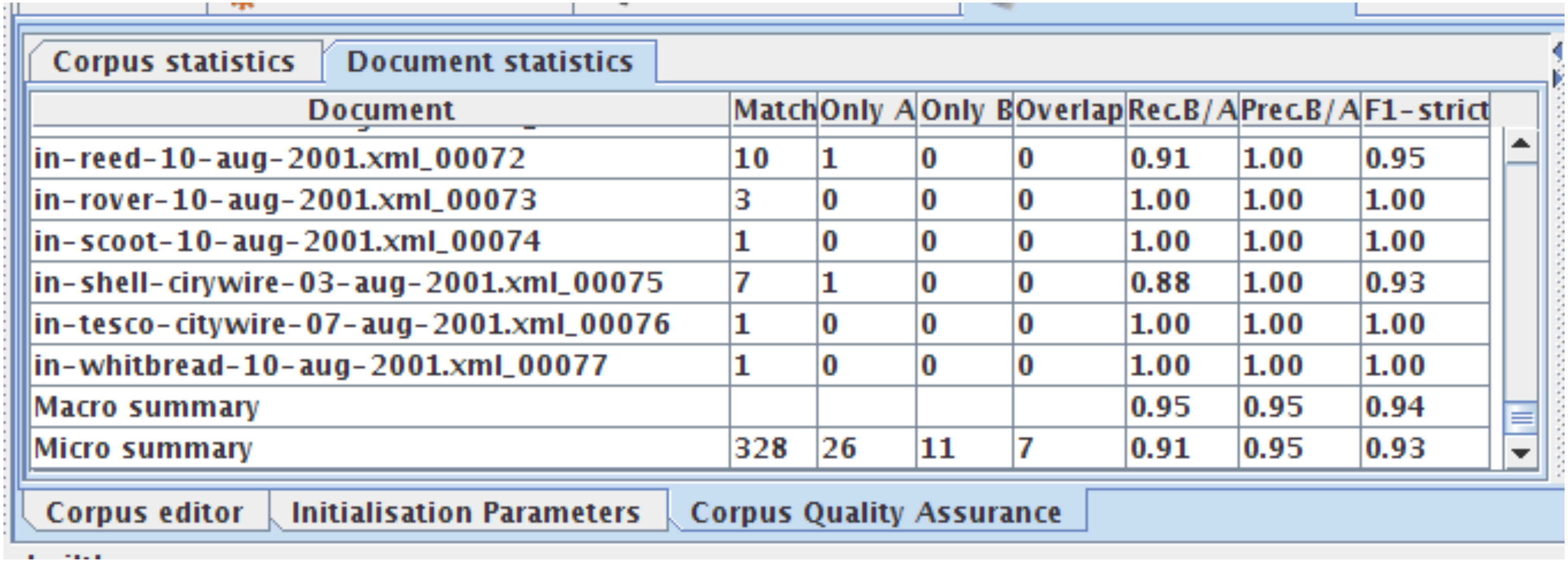
Annotation	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
Person	328	26	11	7	0.91	0.95	0.93
Macro summary					0.91	0.95	0.93
Micro summary	328	26	11	7	0.91	0.95	0.93

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

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

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

Summary

- In this session, we've looked at evaluation for NLP tools, why it's important, and some ways to do it
- Note that for a proper evaluation, the gold standard should ideally be annotated by multiple annotators, and inter-annotator agreement compared
- This is because some of these annotation tasks are quite hard, and you want to be sure that the annotators have done a good job!
- Overall, this module has taught some basic NLP concepts and let you experiment with them in GATE
- You're ready to start building your own tools!

Fun extra task

- If you have time, you can try annotating a document yourself with named entities and then comparing how you did with the existing Key annotation set
- Reminder: to annotate a document, make sure the right annotation set is selected with the mouse (we suggest adding a new one with your name) and then highlight the text you want to annotate. A popup window will appear, letting you select the annotation type.
- Use one of the evaluation tools to compare how you did!