Introduction to GATE

Dr. Paula Matuszek

Paula.Matuszek@gmail.com

Taken partially from a presentation by Lin Lin. http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

What is GATE?

- Stands for General Architecture for Text Engineering.
- Developed at the University of Sheffield
- Component-based architecture with data separated from applications, many discrete capabilities included as plugins.

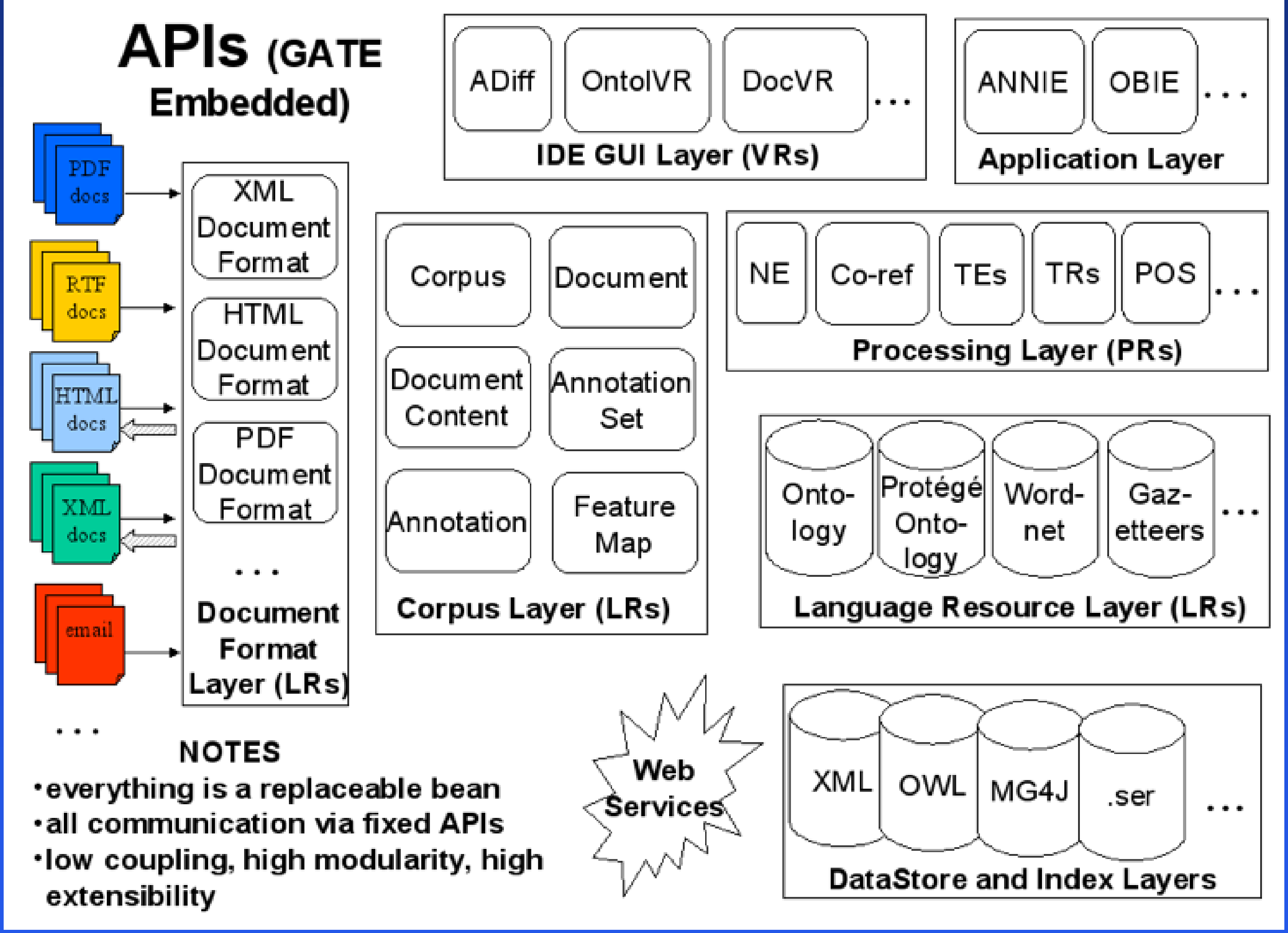
Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Who Uses GATE?

- Scientists performing experiments that involve processing human language
- Developers developing applications with language processing components
- Teachers and students of courses about language and language computation
- Us :-)

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

GATE Architecture Overview



NOTES

- everything is a replaceable bean
- all communication via fixed APIs
- low coupling, high modularity, high extensibility

http://gate.ac.uk/overview.html

GATE Product Family

- GATE Developer: IDE for language processing, with information extraction and other plugins.
- GATE Embedded: object library which can be included in applications
- GATE Teamware: collaborative annotation environment
- GATE Mimir: a “multiparadigm index” which supports semantic indexing and search
- GATE Wiki: “controllable wiki” based on Grails and Subversion
- GATE Cloud: GATE embedded running on supercomputer hardware

GATE Components

- We will deal primarily with GATE Developer:
- It has four components:
  - Applications: groups of processes to be run on a document or corpus.
  - LanguageResources (LRs): entities such as lexicons, documents, corpora, annotation schemas, ontologies.
  - ProcessingResources (PRs): tools that operate on unstructured text, such as parsers and tokenizers. These are mostly plugins.
  - DataStores: saved processed documents and resources.

Overview of Gate Developer

- GATE Developer
- Resources Pane
  - applications: groups of processes to run on a document or corpus
  - language resources: corpus, ontologies, schemas
  - processing resources: tools that operate on unstructured text
  - datastores: saved documents and resources
- Display Pane: whatever you're currently working with.

Language Resources

- Language Resources can be of four kinds:
  - Documents are modeled as content plus annotations plus features.
  - A Corpus is a Java Set whose members are Documents.
  - Annotations are organized in graphs, which are modeled as Java sets of Annotation.
  - Schemas are XML schemas describing allowable annotations and features

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Documents Processing in GATE

- Document:
  - Formats including XML, RTF, email, HTML, SGML, and plain text.
  - Identified and converted into GATE annotation format.
  - Processed by Processing Resources.
  - Results stored in a serial data store (based on Java serialization) or indexed in a Lucene database.
  - Can also be exported as XML.

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

CREOLE

- A Collection of REusable Objects for Language Engineering
- The set of resources integrated with GATE
- All the resources are packaged as Java Archive (or ‘JAR’) files, plus some XML configuration data.
- Managed in the Creole Plugin Manager

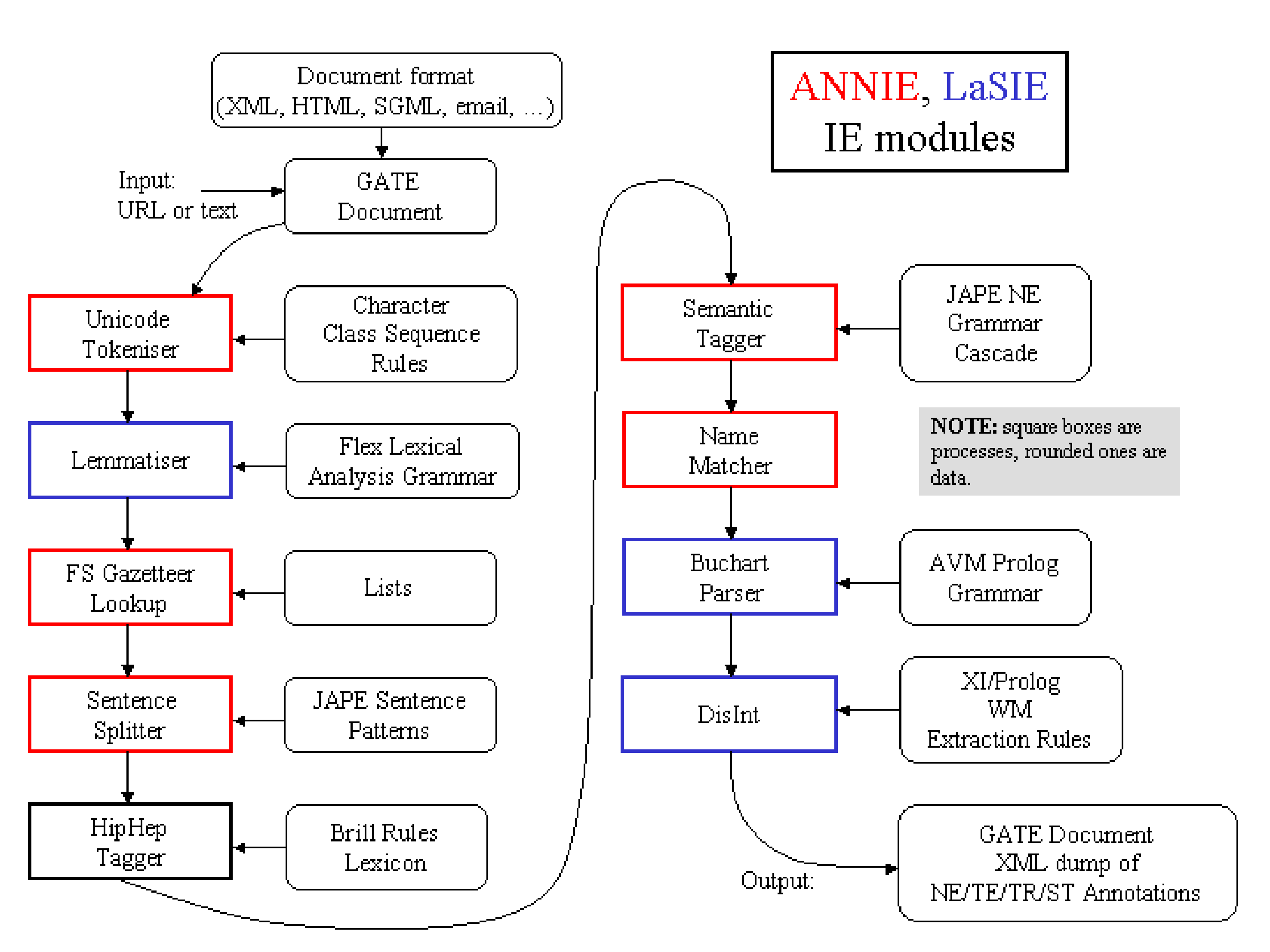
Taken partially from a presentation by Lin Lin. http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Processing Resources: ANNIE

- A family of Processing Resources for language analysis included with GATE
- Stands for A Nearly-New Information Extraction system.
- Using finite state techniques to implement various tasks: tokenization, semantic tagging, verb phrase chunking, and so on.

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

ANNIE IE Modules



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

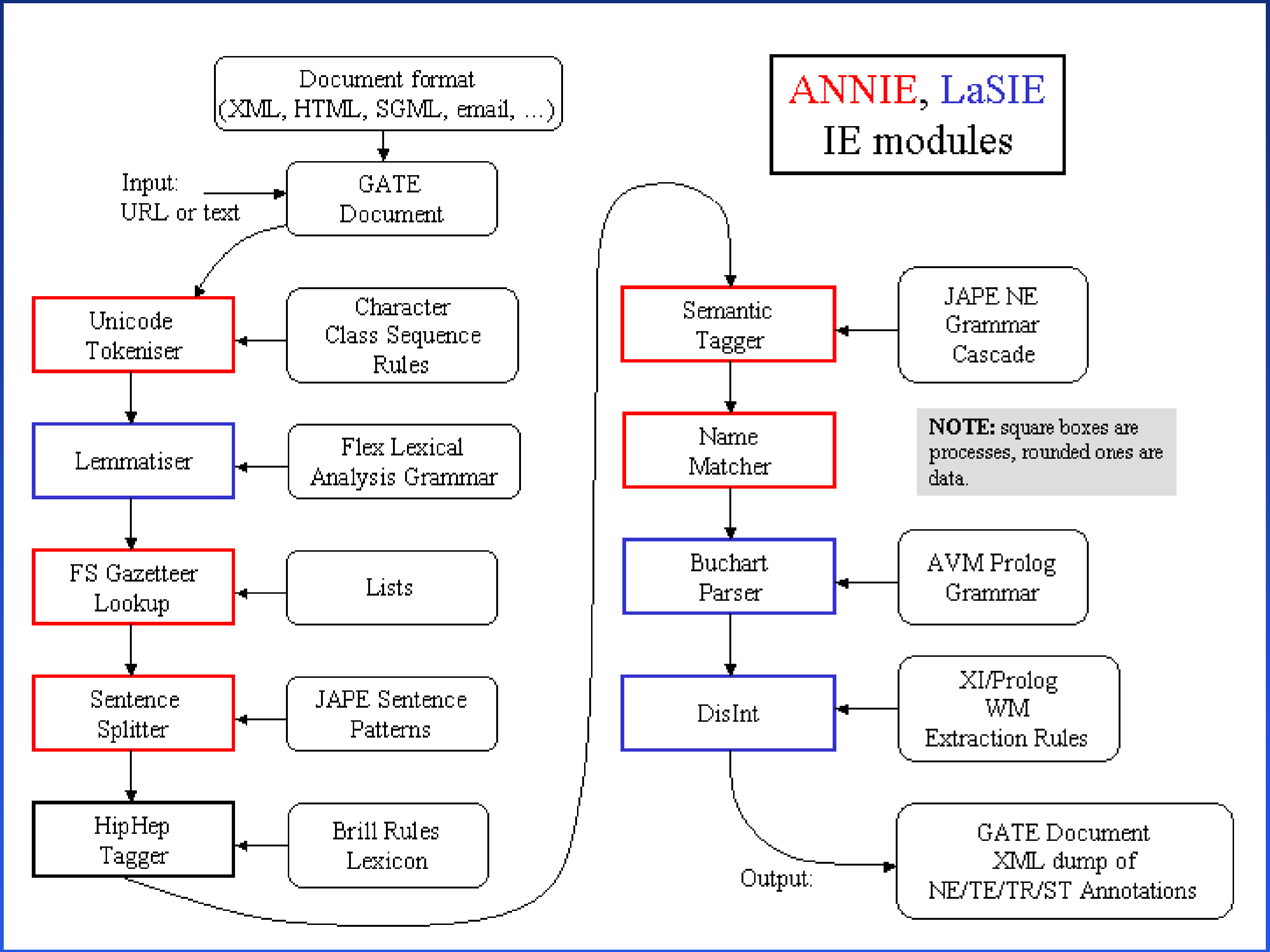
http://gate.ac.uk/sale/tao/splitch6.html#chap:annie

On to ANNIE:

- A family of Processing Resources for language analysis included with GATE
- Stands for A Nearly-New Information Extraction system.
- Using finite state techniques to implement various tasks: tokenization, semantic tagging, verb phrase chunking, and so on.
- (LaSIE is the forerunner of ANNIE, focused specifically on information extraction for the TREC conferences)

GATE and ANNIE Information taken primarily from the GATE user manual, gate.ac.uk/sale/tao, and GATE training materials, http://gate.ac.uk/wiki/training-materials-2011.html

ANNIE IE Modules



ANNIE, LaSIE
IE modules

Document format
(XML, HTML, SGML, email, …)

Input:
URL or text

GATE
Document

Unicode
Tokeniser

Character
Class Sequence
Rules

Lemmatiser

Flex Lexical
Analysis Grammar

FS Gazetteer
Lookup

Lists

Sentence
Splitter

JAPE Sentence
Patterns

HipHep
Tagger

Brill Rules
Lexicon

Semantic
Tagger

JAPE NE
Grammar
Cascade

Name
Matcher

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

Buchart
Parser

AVM Prolog
Grammar

DisInt

XI/Prolog
WM
Extraction Rules

Output:

GATE Document
XML dump of
NE/TE/TR/ST Annotations

GATE and ANNIE Information taken primarily from the GATE user manual, gate.ac.uk/sale/tao, and GATE training materials, http://gate.ac.uk/wiki/training-materials-2011.html

http://gate.ac.uk/sale/tao/splitch6.html#chap:annie

ANNIE Standard Components

- These are what is loaded when you load ANNIE and run the default application
  - Document Reset
  - Tokenizer
  - Gazetteer: lists of entities
  - Sentence Splitter/Regex sentence splitter
  - Part of Speech Tagger
  - Named Entity Transducer
  - Orthomatcher

GATE and ANNIE Information taken primarily from the GATE user manual, gate.ac.uk/sale/tao, and GATE training materials, http://gate.ac.uk/wiki/training-materials-2011.html

Create an Application with Processing Resources (PRs)

- Applications model a control strategy for the execution of PRs.
- Simple pipelines: group a set of PRs together in order and execute them in turn.
- Corpus pipelines: open each document in the corpus in turn, set that document as a runtime parameter on each PR, run all the PRs on the corpus, then close the document
- We will do this during lab.

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Saving GATE Language Resources and Applications

- Data Stores:
  - save processed documents for additional use
    - specialized folder on a hard drive
    - Lucene database
  - improve processing times for large collections of documents

Taken partially from a presentation by Lin Lin. http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Types of Data Store

- Serial Data Store:
  - based on java’s serialization system.
  - store in a directory
- Lucene Data Store (Lucene is an open-source indexing and search tool.)
  - searchable repository
  - Lucene-based indexing

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Saving in a datastore

- Create a folder.
- Right-click to get Create Datastore menu
- This only creates the store. Save corpora or documents in the Language Resources pane.
- Once saved, they can be

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Saving as XML

- Individual documents can also be saved directly.
  - Special GATE XML format
    - annotations are appended to the document, locations for tags are embedded in body
  - Preserve original format
    - use for XML or html.
    - will save all original tags and everything selected in the annotations
    - For a plain text file, embeds inline tags.

Taken partially from a presentation by Lin Lin.
http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

Saving Applications

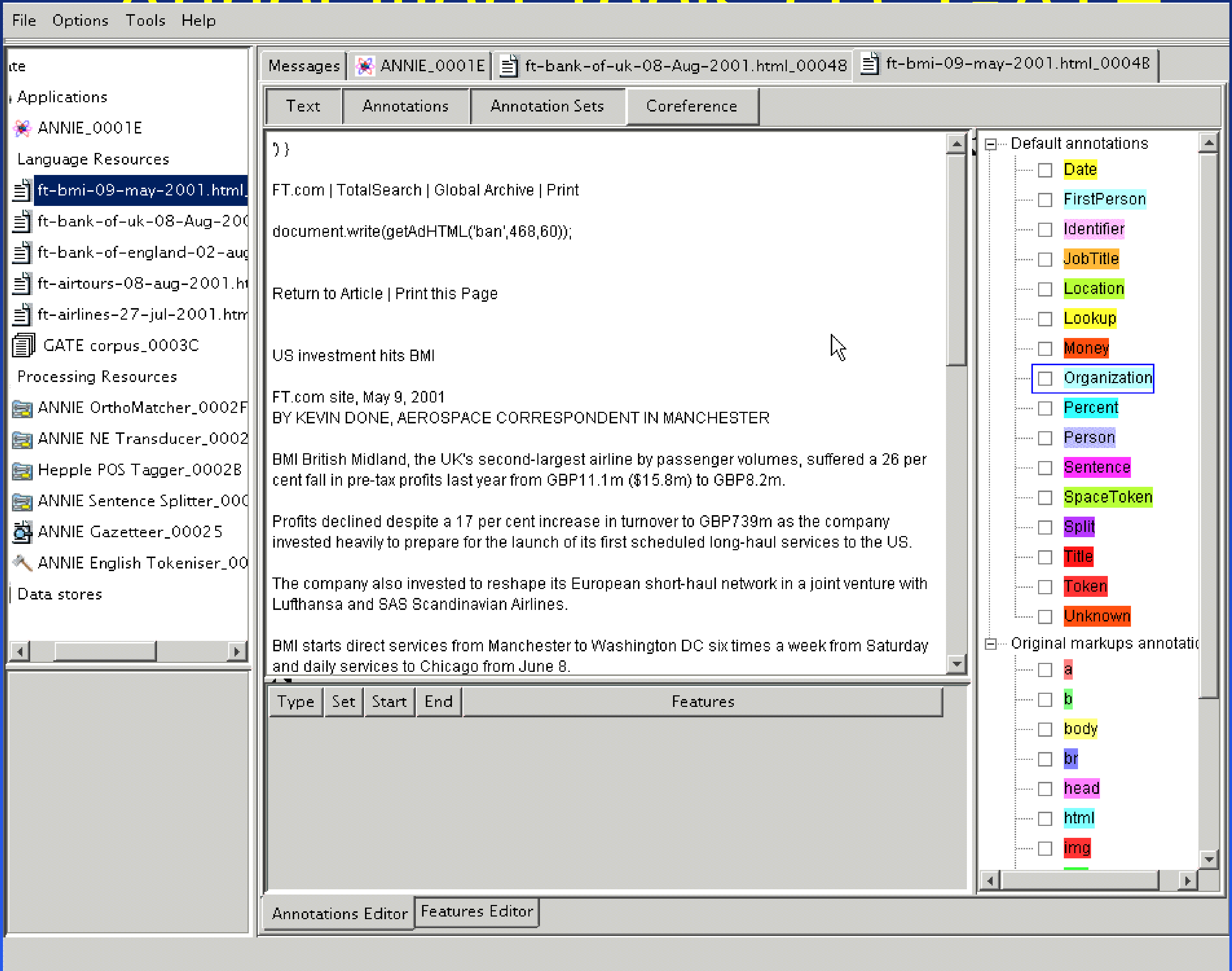
- Save a set of processing resources and their parameters.
  - Right-click, save application state.
  - Append .xgapp for name
- To export as a standalone, export as teamware
  - bundles all needed files
  - intended for teamware but can be used for sharing directly.

Taken partially from a presentation by Lin Lin. http://iwayan.info/Research/Interoperability/Tutor_Workshop/AmitShethGlobalInfInfrastucture/Presentation/GATE.ppt

And LOTS more

- GATE is an extraordinarily rich system. Some of the other CREOLE resources included in the standard distribution:
  - Annotation Merging, Quality assurance summarizer for comparing annotations
  - Web crawler , Information Retrieval, Key Phrase Extraction
  - Machine learning
  - Domain-specific taggers (e.g., chemistry)
  - Resources for many languages
- CREOLE plugins for integrating with many other systems. E.g.
  - UIMA
  - Wordnet
  - Penn BioTagger
  - OpenCalais
  - OpenNLP
  - LingPipe
- More details at http://gate.ac.uk/gate/doc/plugins.html

Annotation Tools (1): GATE



') }

FT.com | TotalSearch | Global Archive | Print

document.write(getAdHTML('ban',468,60));

Return to Article | Print this Page

US investment hits BMI

FT.com site, May 9, 2001
BY KEVIN DONE, AEROSPACE CORRESPONDENT IN MANCHESTER

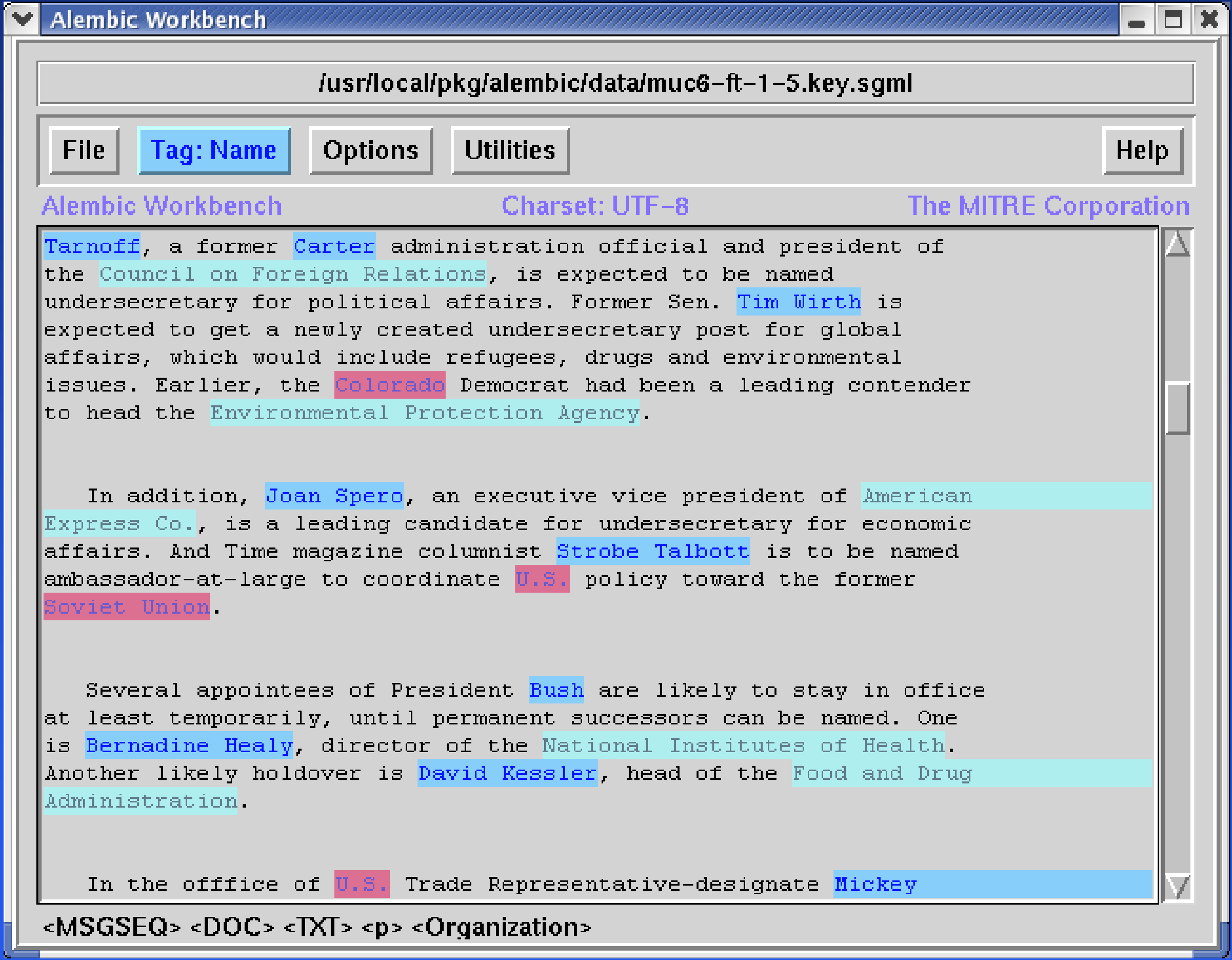
BMI British Midland, the UK's second-largest airline by passenger volumes, suffered a 26 per cent fall in pre-tax profits last year from GBP11.1m ($15.8m) to GBP8.2m.

Profits declined despite a 17 per cent increase in turnover to GBP739m as the company invested heavily to prepare for the launch of its first scheduled long-haul services to the US.

The company also invested to reshape its European short-haul network in a joint venture with Lufthansa and SAS Scandinavian Airlines.

BMI starts direct services from Manchester to Washington DC six times a week from Saturday and daily services to Chicago from June 8.

Annotation Tools (2): Alembic



Tarnoff, a former Carter administration official and president of the Council on Foreign Relations, is expected to be named undersecretary for political affairs. Former Sen. Tim Wirth is expected to get a newly created undersecretary post for global affairs, which would include refugees, drugs and environmental issues. Earlier, the Colorado Democrat had been a leading contender to head the Environmental Protection Agency.

In addition, Joan Spero, an executive vice president of American Express Co., is a leading candidate for undersecretary for economic affairs. And Time magazine columnist Strobe Talbott is to be named ambassador-at-large to coordinate U.S. policy toward the former Soviet Union.

Several appointees of President Bush are likely to stay in office at least temporarily, until permanent successors can be named. One is Bernadine Healy, director of the National Institutes of Health. Another likely holdover is David Kessler, head of the Food and Drug Administration.

In the offfice of U.S. Trade Representative-designate Mickey

<MSGSEQ> <DOC> <TXT> <p> <Organization>

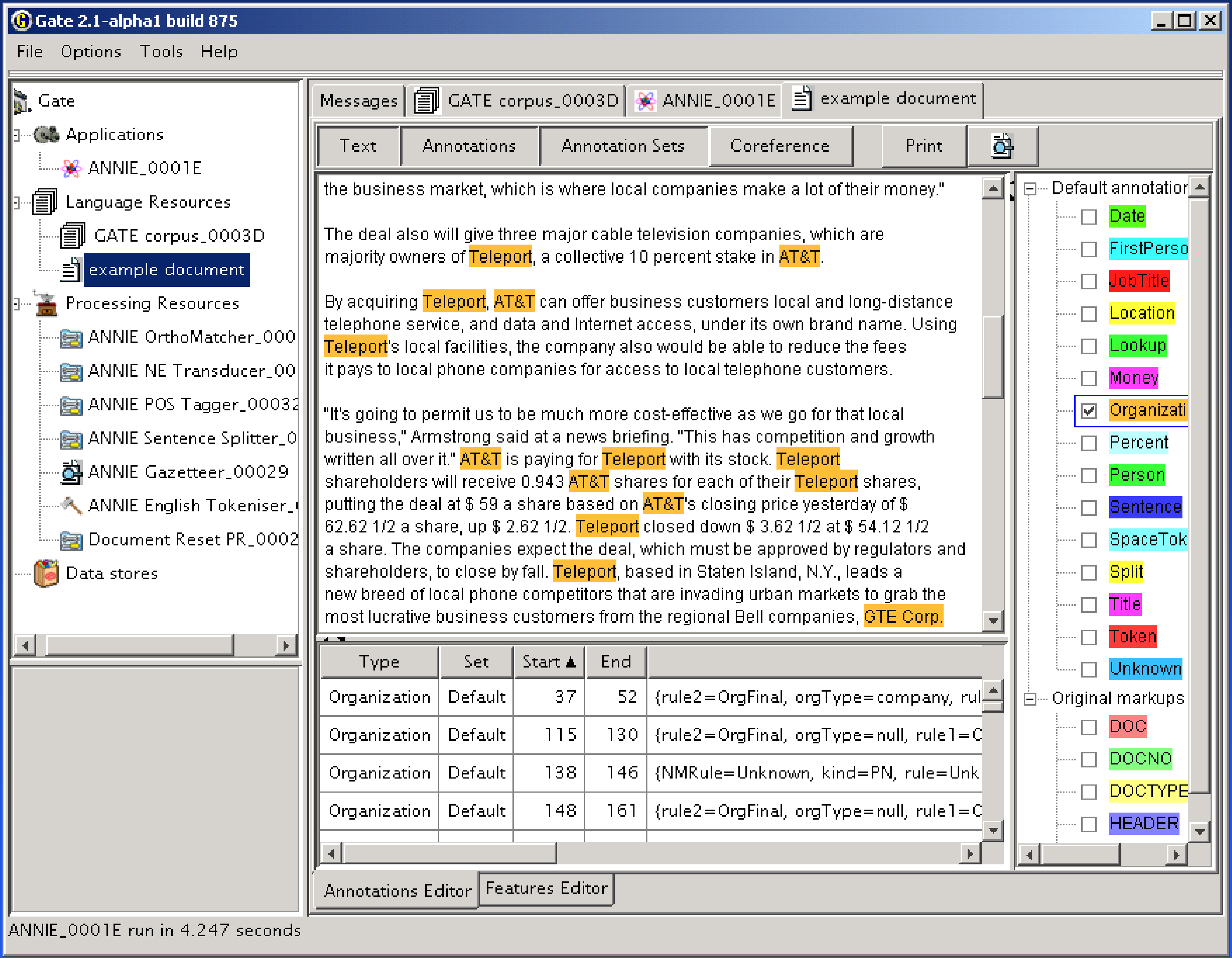
NE Rule in JAPE

JAPE: a Java Annotation Patterns Engine

- Light, robust regular-expression-based processing
- Cascaded finite state transduction
- Low-overhead development of new components
- Simplifies multi-phase regex processing

```
Rule: Company1
Priority: 25
   (
     ( {Token.orthography == upperInitial} )+ //from tokeniser
     {Lookup.kind == companyDesignator} //from gazetteer lists
   ):match
-->
    :match.NamedEntity =
      { kind=company, rule=“Company1” }
```

Named Entities in GATE



the business market, which is where local companies make a lot of their money."

The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T.

By acquiring Teleport, AT&T can offer business customers local and long-distance telephone service, and data and Internet access, under its own brand name. Using Teleport's local facilities, the company also would be able to reduce the fees it pays to local phone companies for access to local telephone customers.

"It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at $ 59 a share based on AT&T's closing price yesterday of $ 62.62 1/2 a share, up $ 2.62 1/2. Teleport closed down $ 3.62 1/2 at $ 54.12 1/2 a share. The companies expect the deal, which must be approved by regulators and shareholders, to close by fall. Teleport, based in Staten Island, N.Y., leads a new breed of local phone competitors that are invading urban markets to grab the most lucrative business customers from the regional Bell companies, GTE Corp.

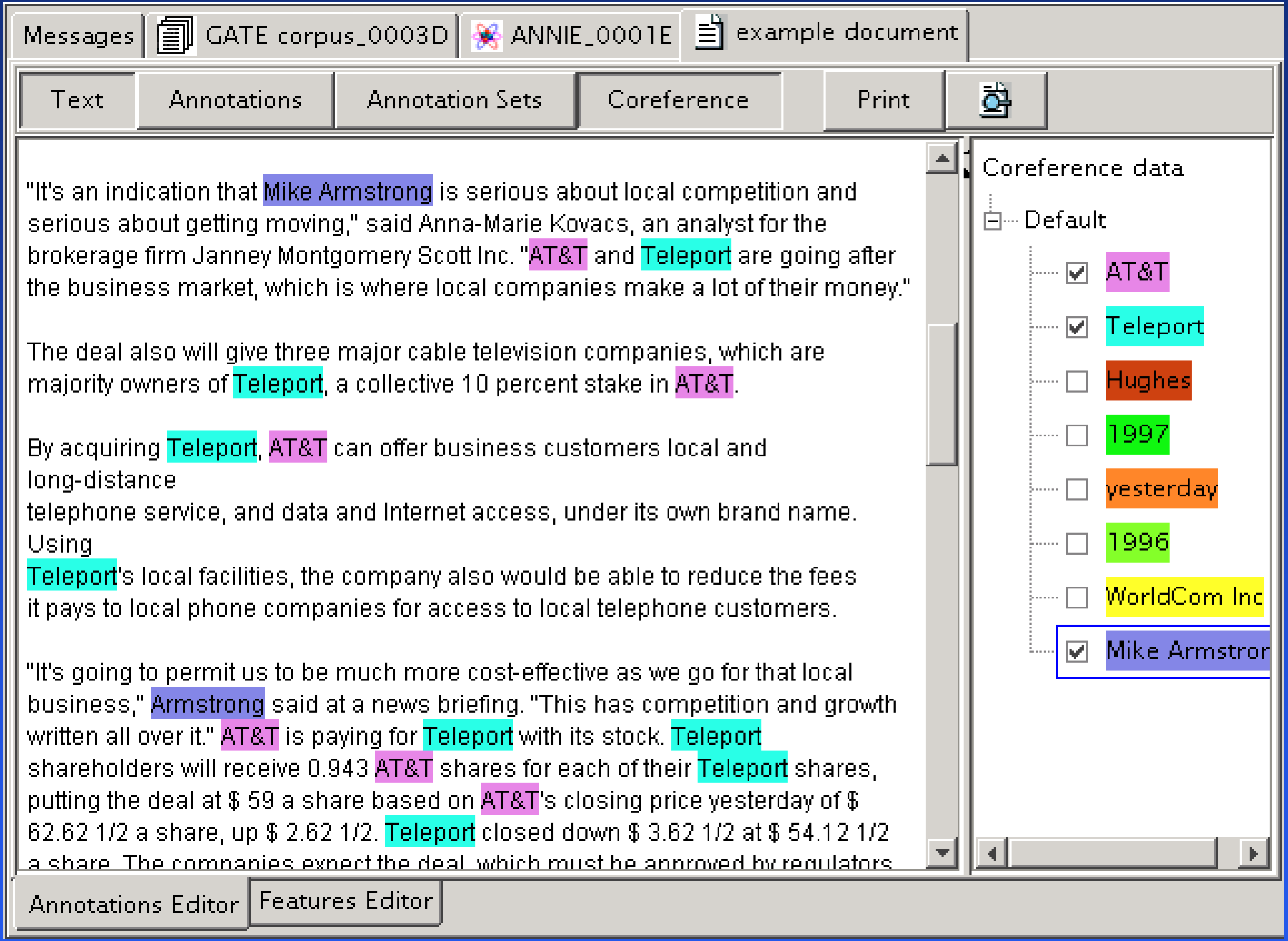
Default annotatior
- □ Date
- □ FirstPerso
- □ JobTitle
- □ Location
- □ Lookup
- □ Money
- ☑ Organizati
- □ Percent
- □ Person
- □ Sentence
- □ SpaceTok
- □ Split
- □ Title
- □ Token
- □ Unknown

Original markups
- □ DOC
- □ DOCNO
- □ DOCTYPE
- □ HEADER

Type	Set	Start ▲	End	
Organization	Default	37	52	{rule2=OrgFinal, orgType=company, rul
Organization	Default	115	130	{rule2=OrgFinal, orgType=null, rule1=C
Organization	Default	138	146	{NMRule=Unknown, kind=PN, rule=Unk
Organization	Default	148	161	{rule2=OrgFinal, orgType=null, rule1=C

ANNIE_0001E run in 4.247 seconds

Named Entity Coreference



"It's an indication that Mike Armstrong is serious about local competition and serious about getting moving," said Anna-Marie Kovacs, an analyst for the brokerage firm Janney Montgomery Scott Inc. "AT&T and Teleport are going after the business market, which is where local companies make a lot of their money."

The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T.

By acquiring Teleport, AT&T can offer business customers local and
long-distance
telephone service, and data and Internet access, under its own brand name.
Using
Teleport's local facilities, the company also would be able to reduce the fees
it pays to local phone companies for access to local telephone customers.

"It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at $ 59 a share based on AT&T's closing price yesterday of $ 62.62 1/2 a share, up $ 2.62 1/2. Teleport closed down $ 3.62 1/2 at $ 54.12 1/2 a share. The companies expect the deal, which must be approved by regulators

Coreference data

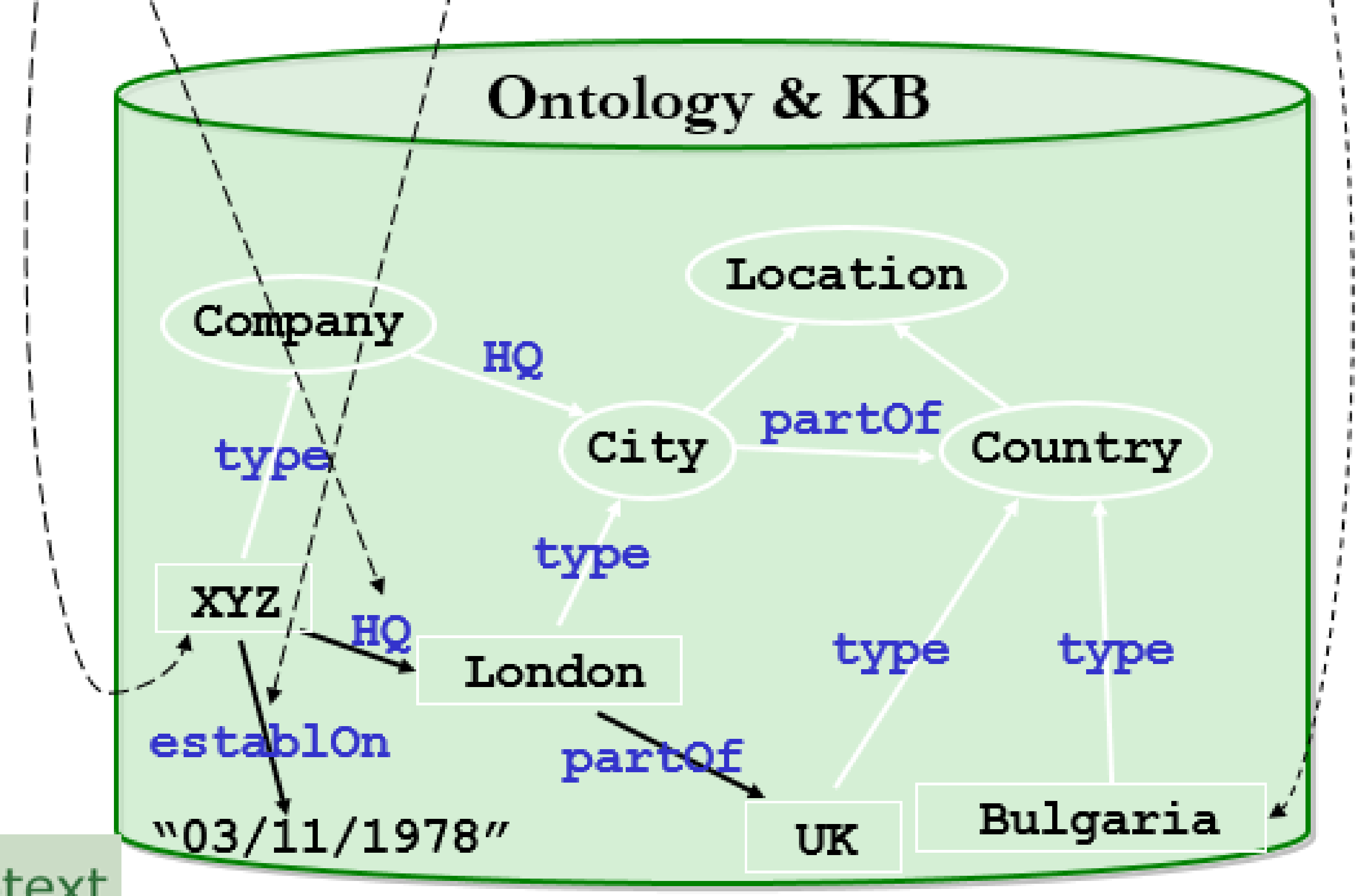
Default
- ☑ AT&T
- ☑ Teleport
- ☐ Hughes
- ☐ 1997
- ☐ yesterday
- ☐ 1996
- ☐ WorldCom Inc
- ☑ Mike Armstror

Semantic Mapping to Ontologies

- Identify entity mentions in the text
- Reference disambiguation
  - Add new instances if needed
  - Disambiguate wrt instances in the ontology
- Identify instances of attributes and relations
  - take into account what are allowed given the **ontology**, using domain&range as constraints

Example

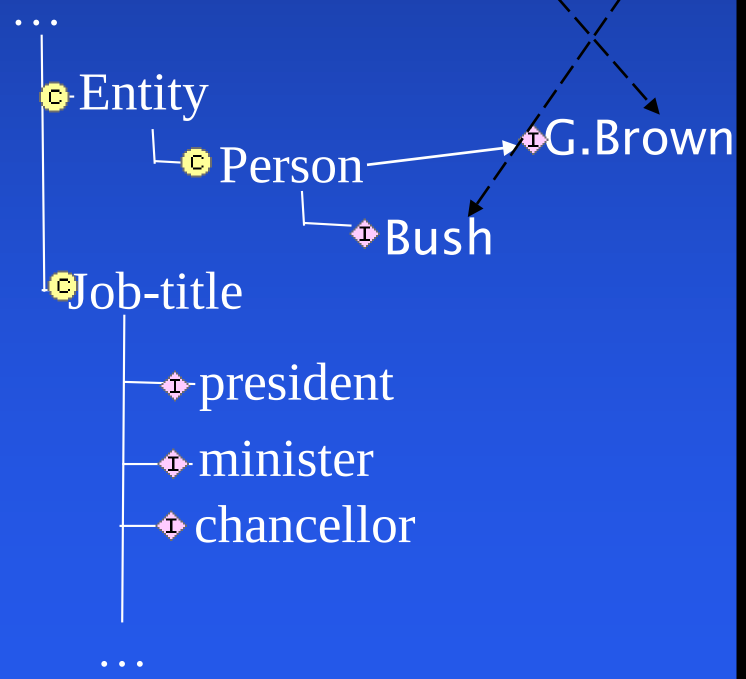
XYZ was established on 03 November 1978 in London. It opened a plant in Bulgaria in …



Classes, instances & metadata

“Gordon Brown met George Bush during his two day visit.

Classes+instances before after



```
<metadata>
  <DOC-ID>http://… 1.html</DOC-ID>
  <Annotation>
    <s_offset> 0 </s_offset>
    <e_offset> 12 </e_offset>
    <string>Gordon Brown</string>
    <class>…#Person</class>
    <inst>…#Person12345</inst>
  </Annotation>
 <Annotation>
    <s_offset> 18 </s_offset>
    <e_offset> 32 </e_offset>
    <string>George Bush</string>
    <class>…#Person</class>
    <inst>…#Person67890</inst>
  </Annotation>
</metadata>
```