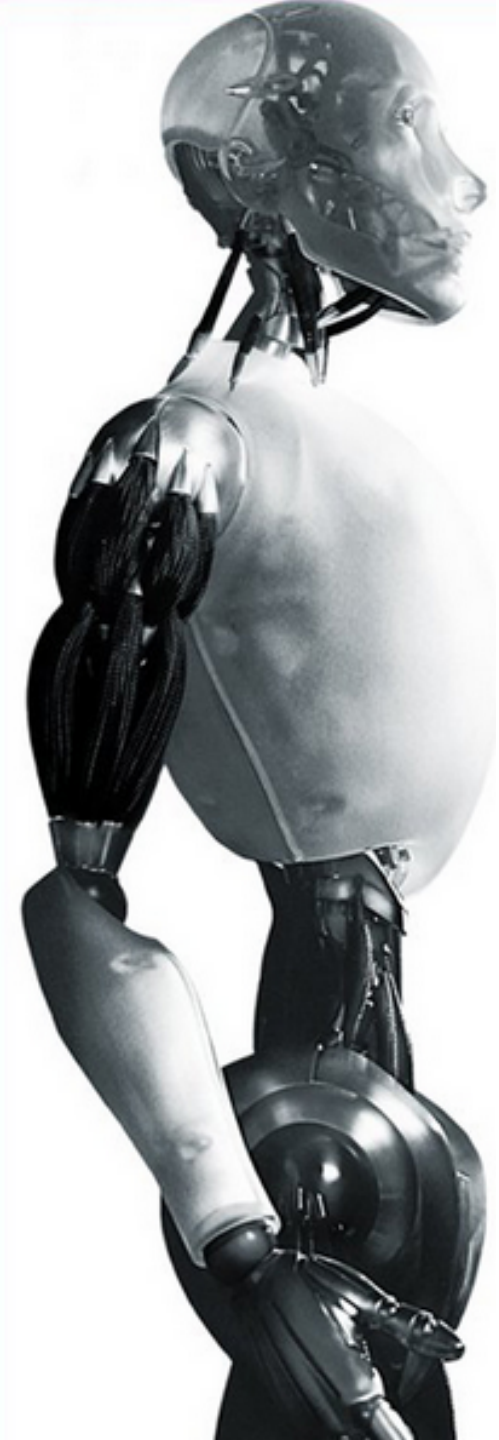
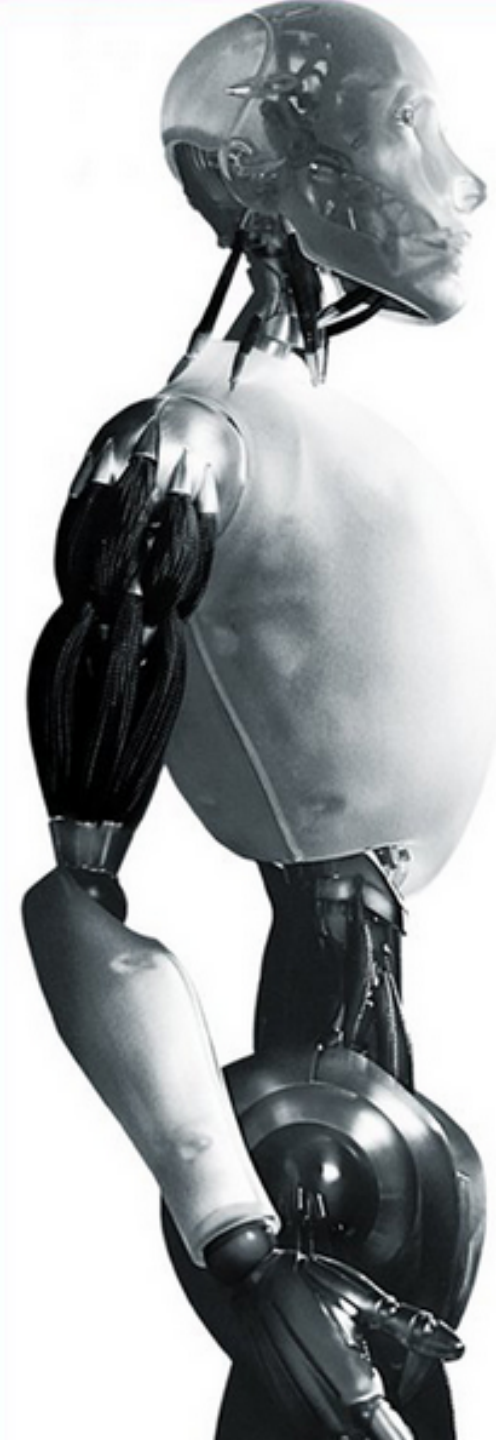


Artificial Intelligence

GATE Case Study

Dr Alexiei Dingli



GATE and ANNIE

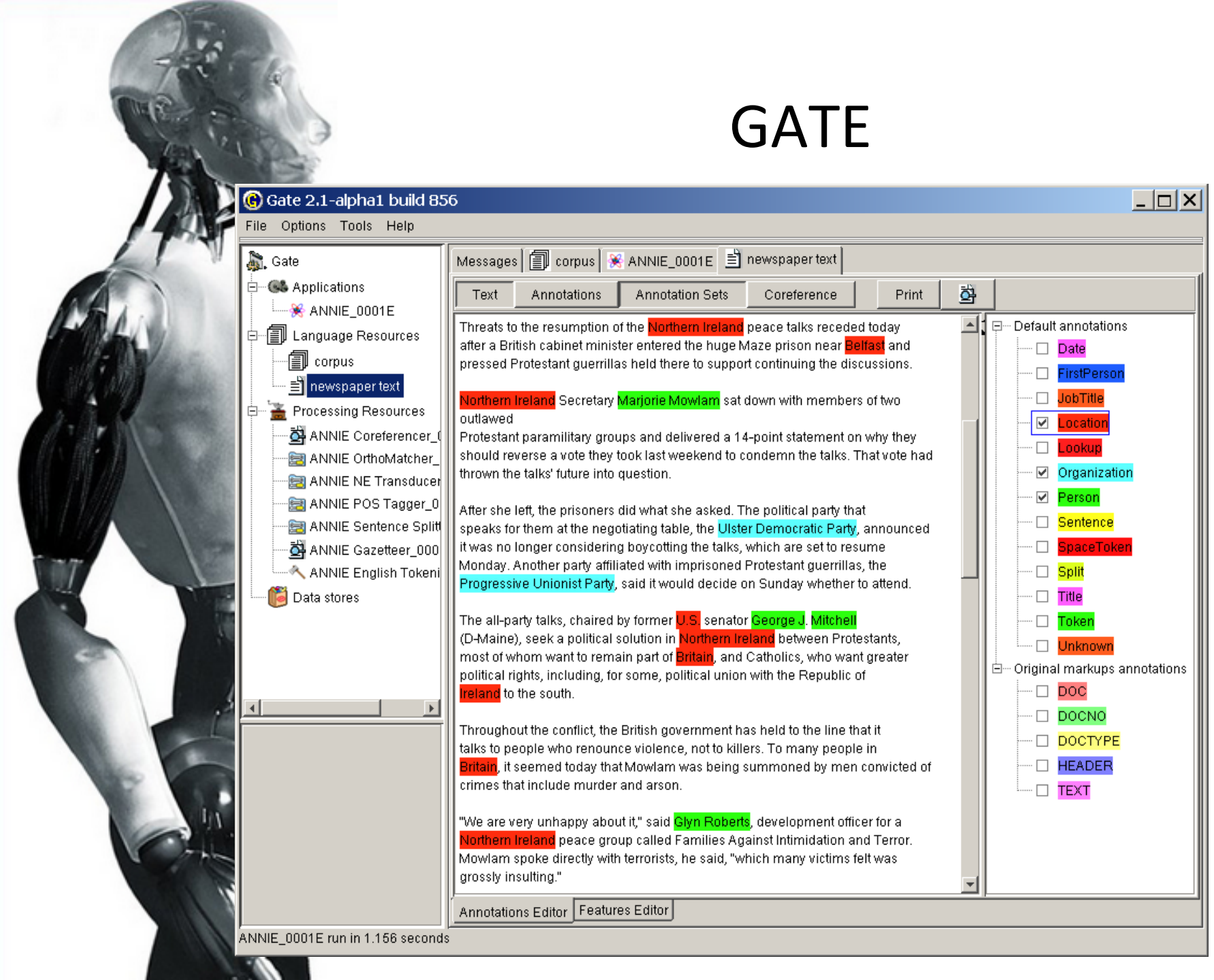
- GATE (Generalised Architecture for Text Engineering) is a framework for language processing

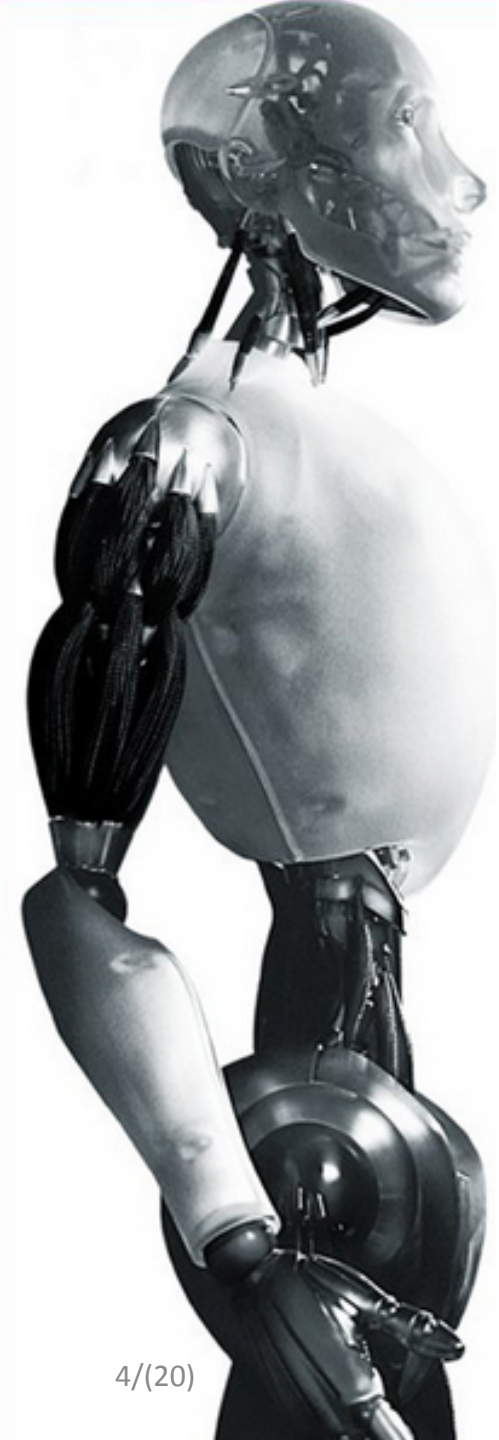
- ANNIE (A Nearly New Information Extraction system) is a suite of language processing tools, which provides NE recognition

GATE also includes:

- plugins for language processing, e.g. parsers, machine learning tools, stemmers, IR tools, IE components for various languages etc.
- tools for visualising and manipulating ontologies
- ontology-based information extraction tools
- evaluation and benchmarking tools

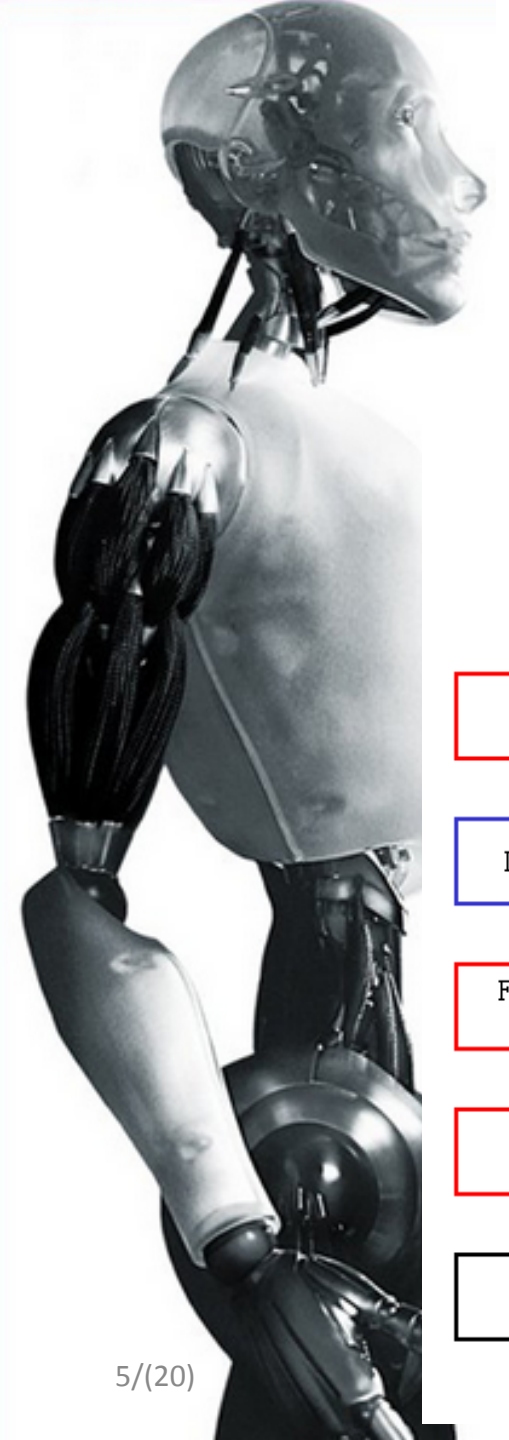
GATE


Gate 2.1-alpha1 build 856
File Options Tools Help
Gate
Applications
ANNIE_0001E
Language Resources
corpus
newspaper text
Processing Resources
ANNIE Coreferencer_(
ANNIE OrthoMatcher_
ANNIE NE Transducer
ANNIE POS Tagger_0
ANNIE Sentence Splitt
ANNIE Gazetteer_000
ANNIE English Tokeni
Data stores
Messages
corpus
ANNIE_0001E
newspaper text
Text
Annotations
Annotation Sets
Coreference
Print
Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.
Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed
Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.
After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.
The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.
Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.
"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."
Default annotations
Date
FirstPerson
JobTitle
Location
Lookup
Organization
Person
Sentence
SpaceToken
Split
Title
Token
Unknown
Original markups annotations
DOC
DOCNO
DOCTYPE
HEADER
TEXT
Annotations Editor
Features Editor
ANNIE_0001E run in 1.156 seconds

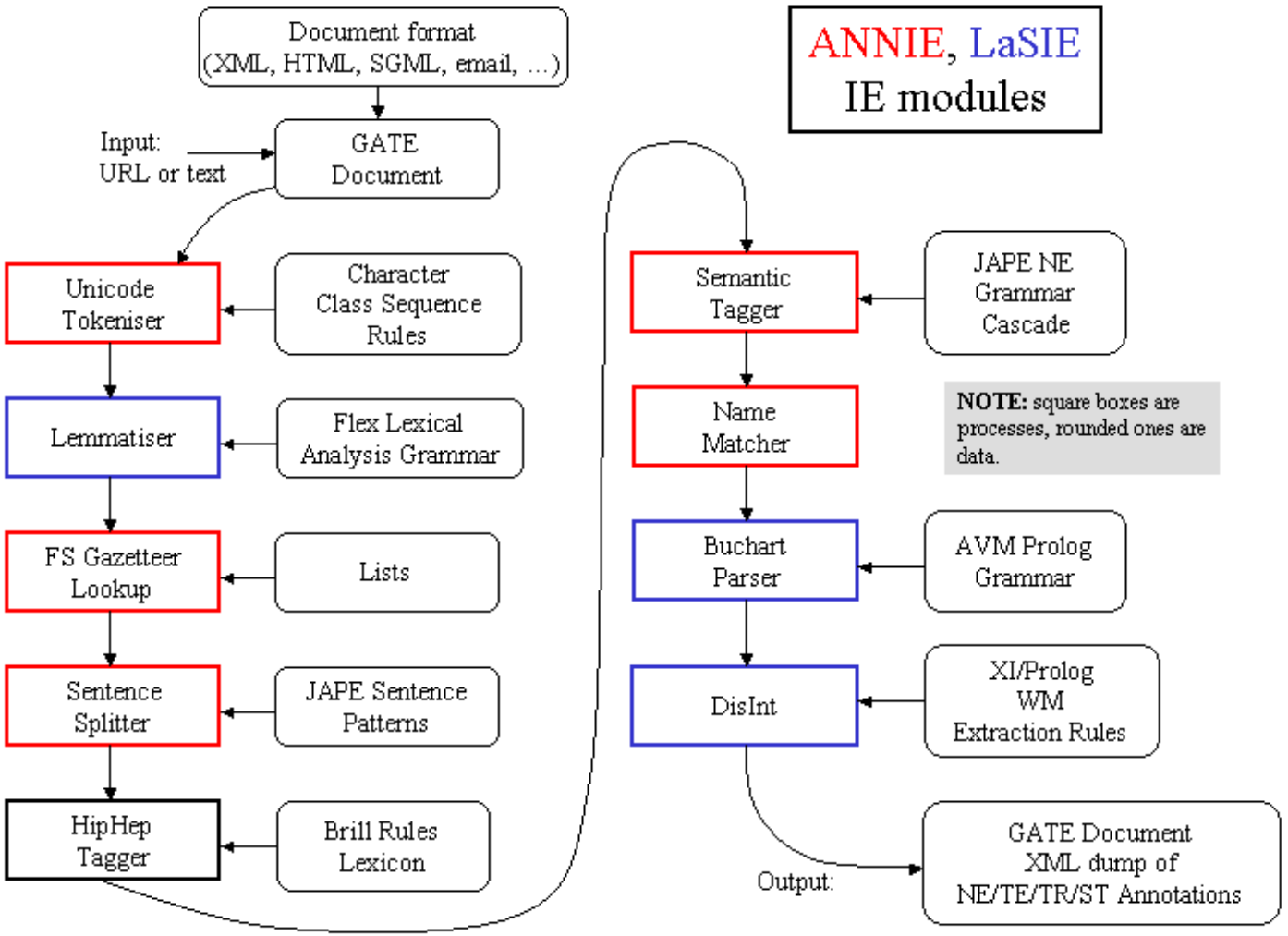


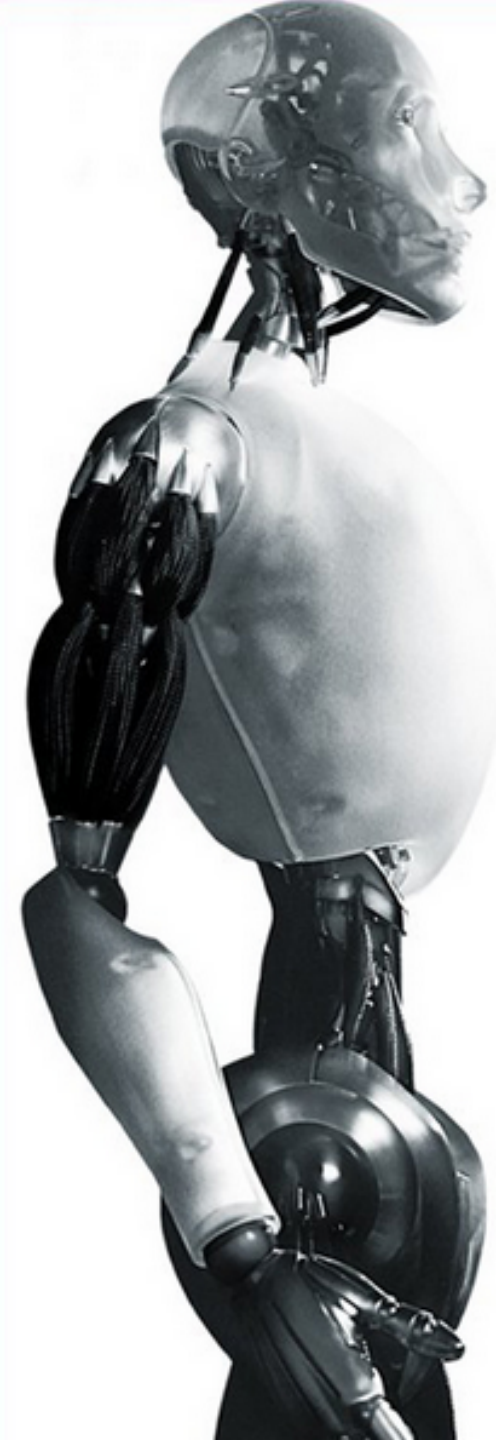
What is ANNIE?

- ANNIE is a vanilla information extraction system comprising a set of core PRs:
  - Tokeniser
  - Sentence Splitter
  - POS tagger
  - Gazetteers
  - Semantic tagger (JAPE transducer)
  - Orthomatcher (orthographic coreference)



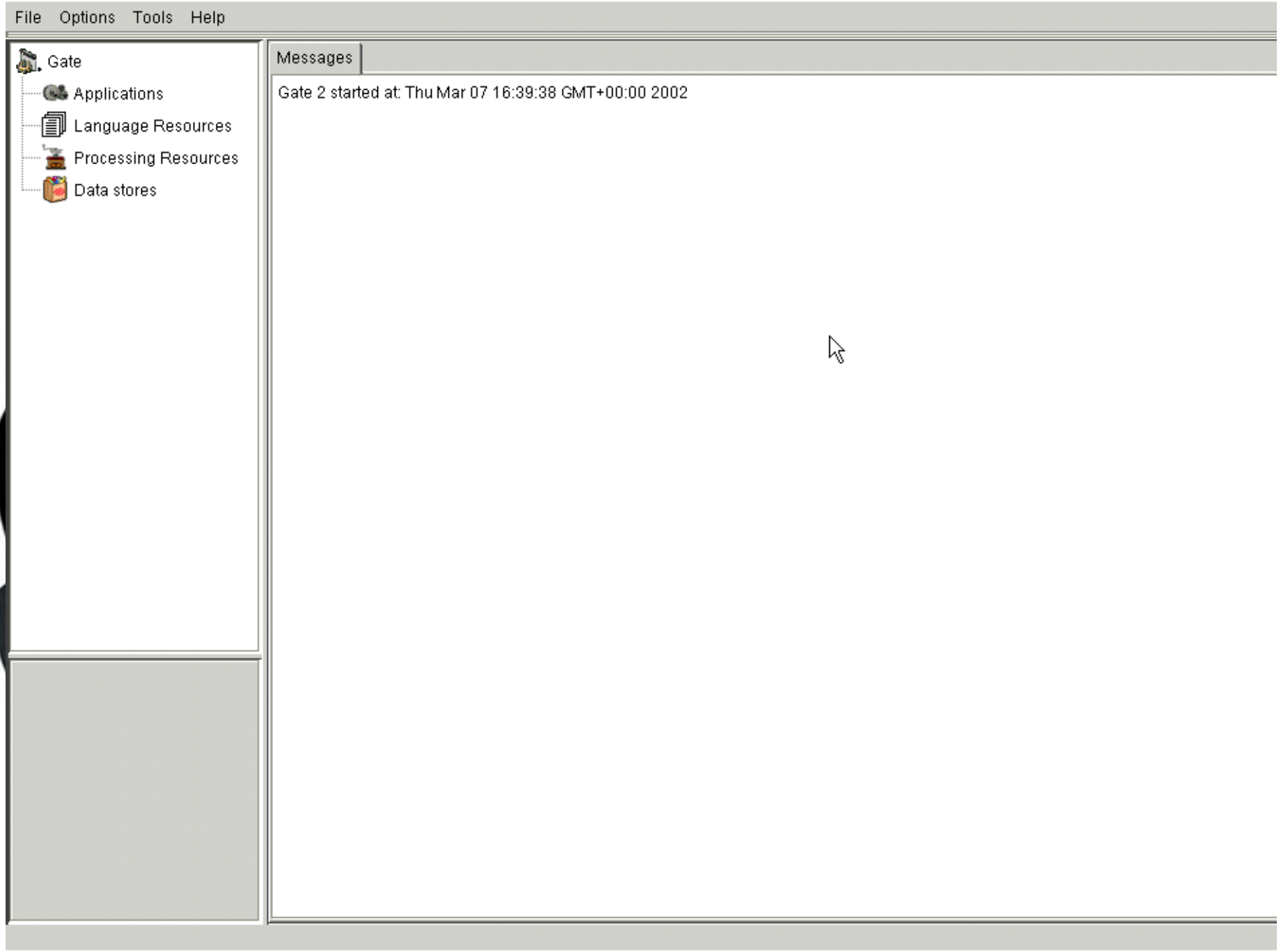
ANNIE Pipeline

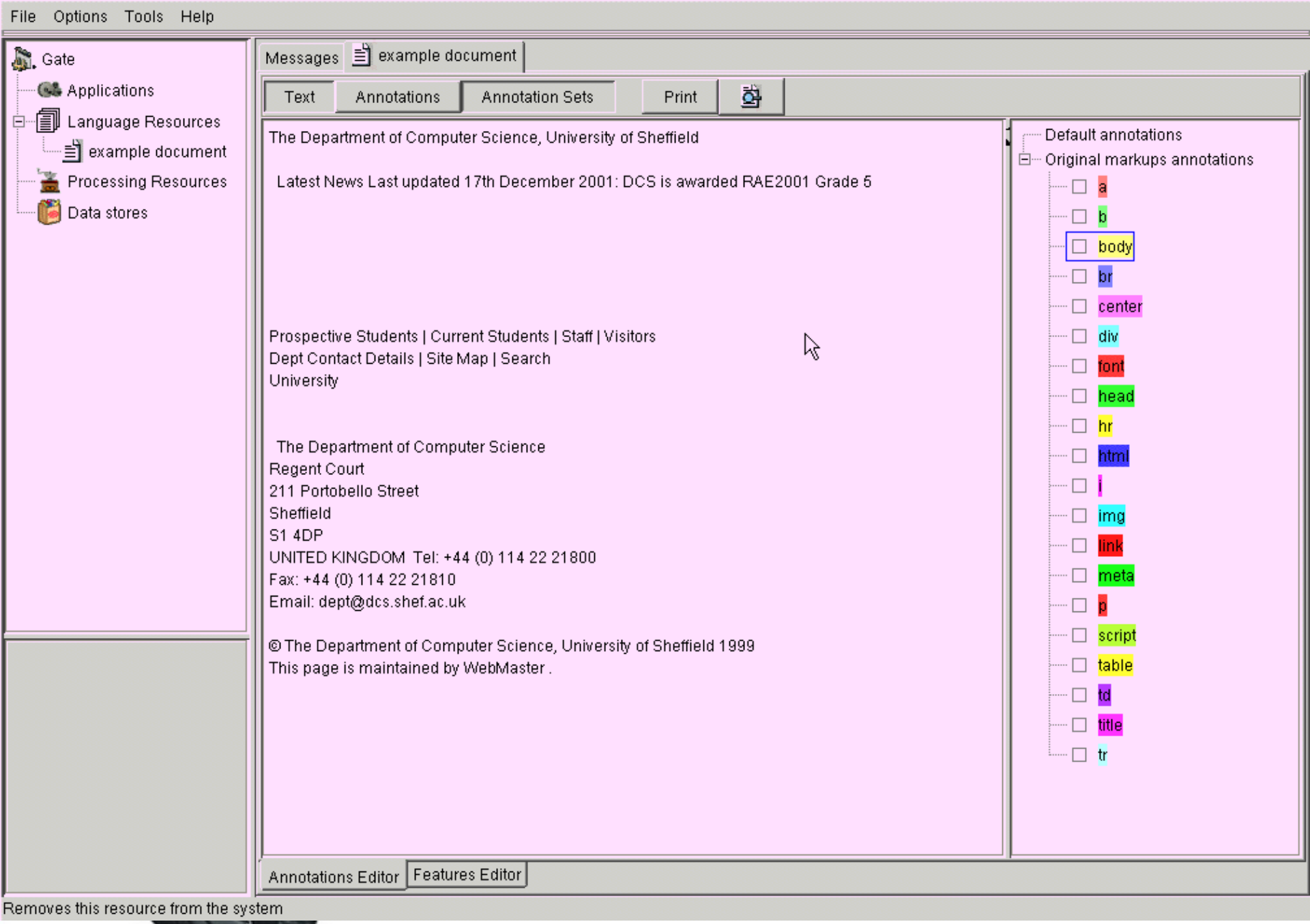


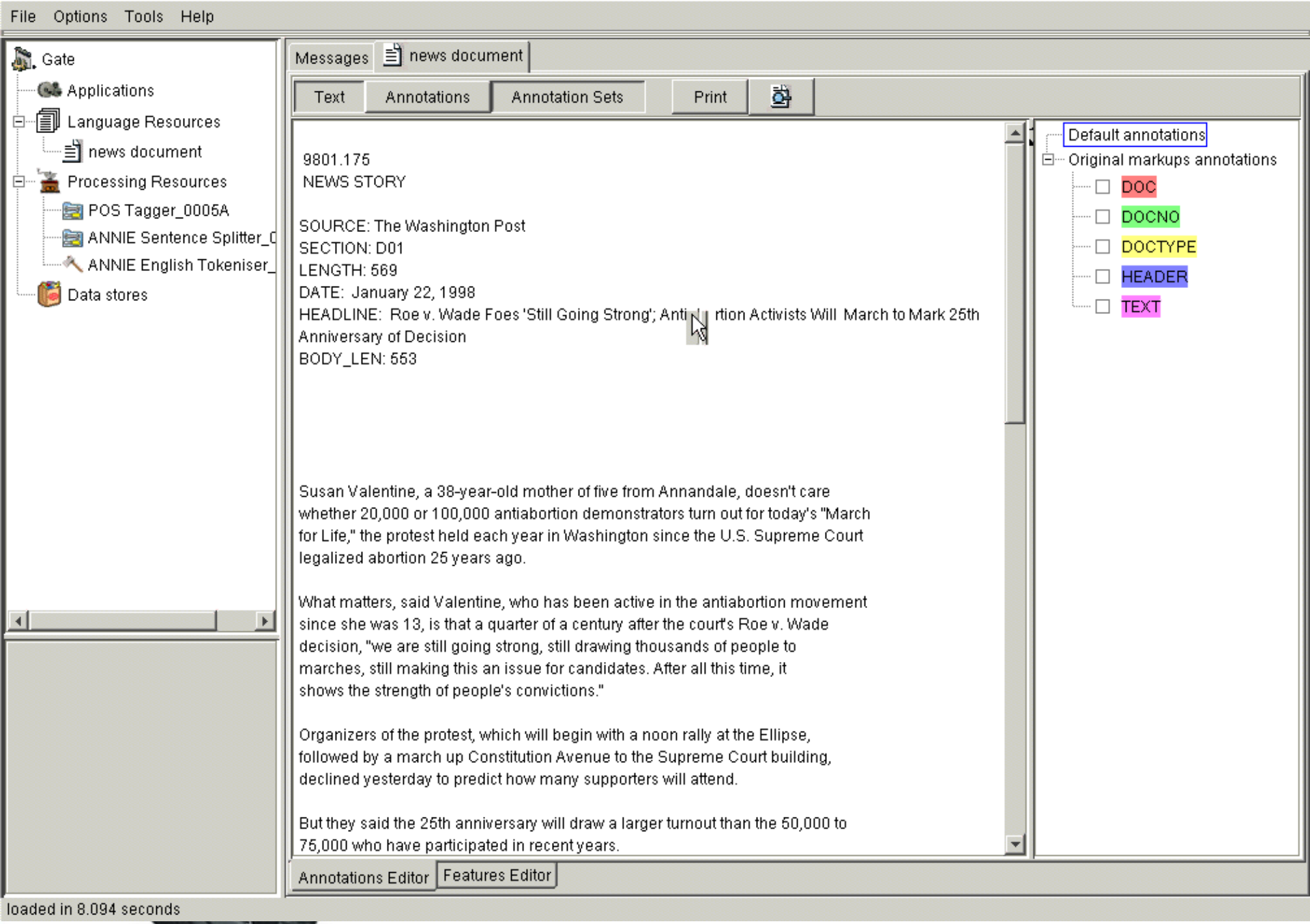


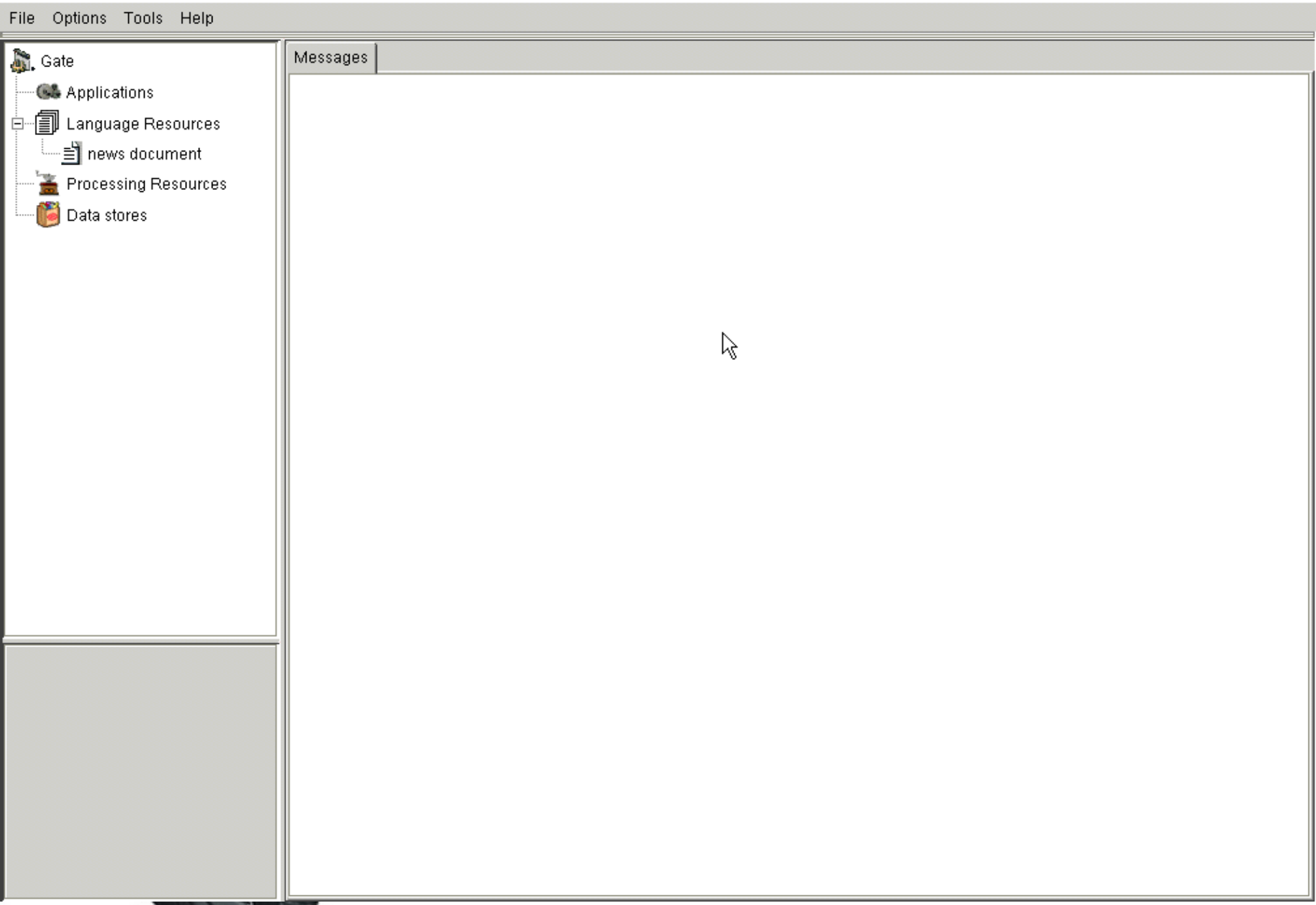
We need a corpus

- Download …
- http://www.cs.cmu.edu/~dayne/SeminarAnnouncements/__Source__.html

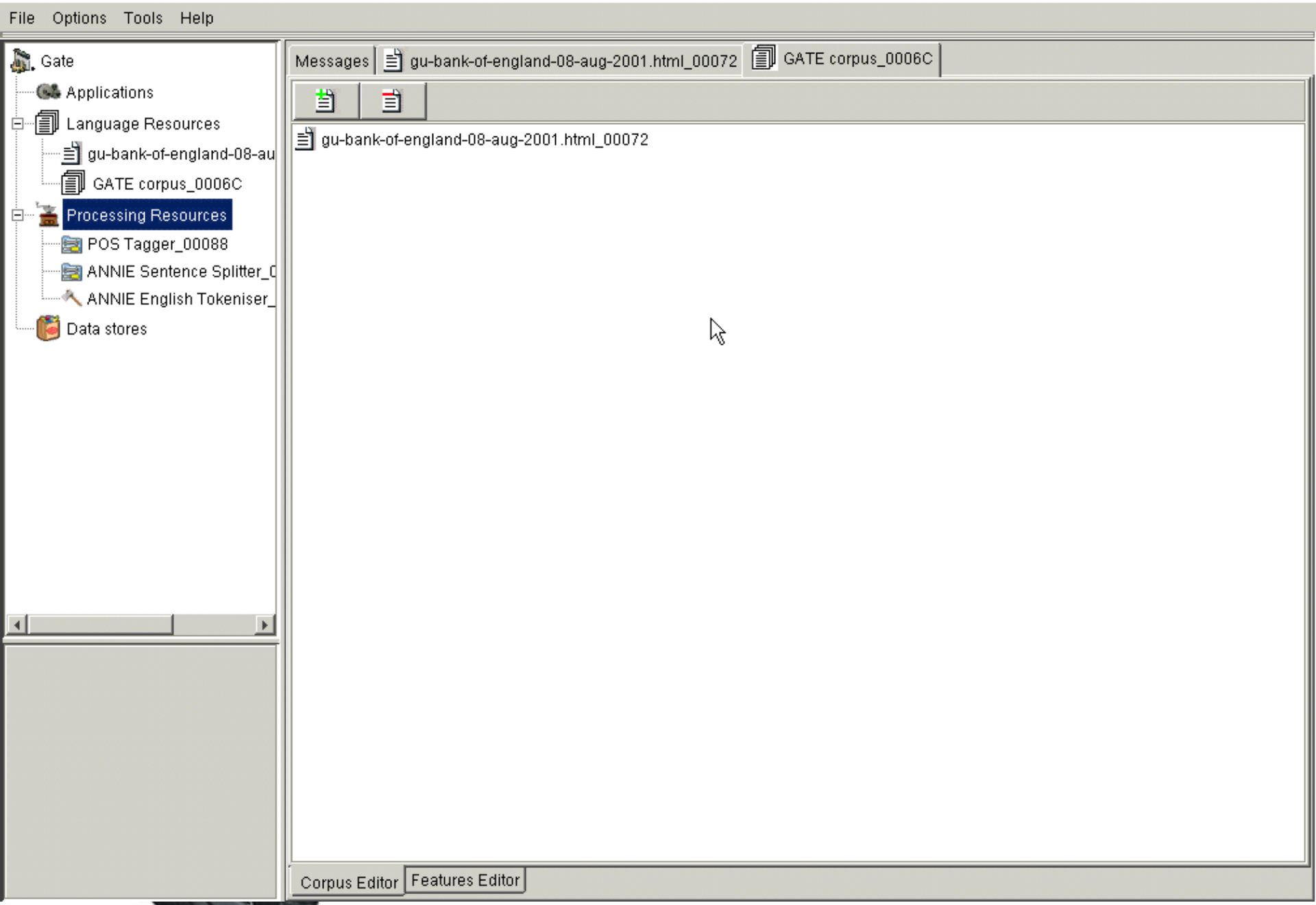

File
Options
Tools
Help
Gate
Applications
Language Resources
Processing Resources
Data stores
Messages
Gate 2 started at: Thu Mar 07 16:39:38 GMT+00:00 2002

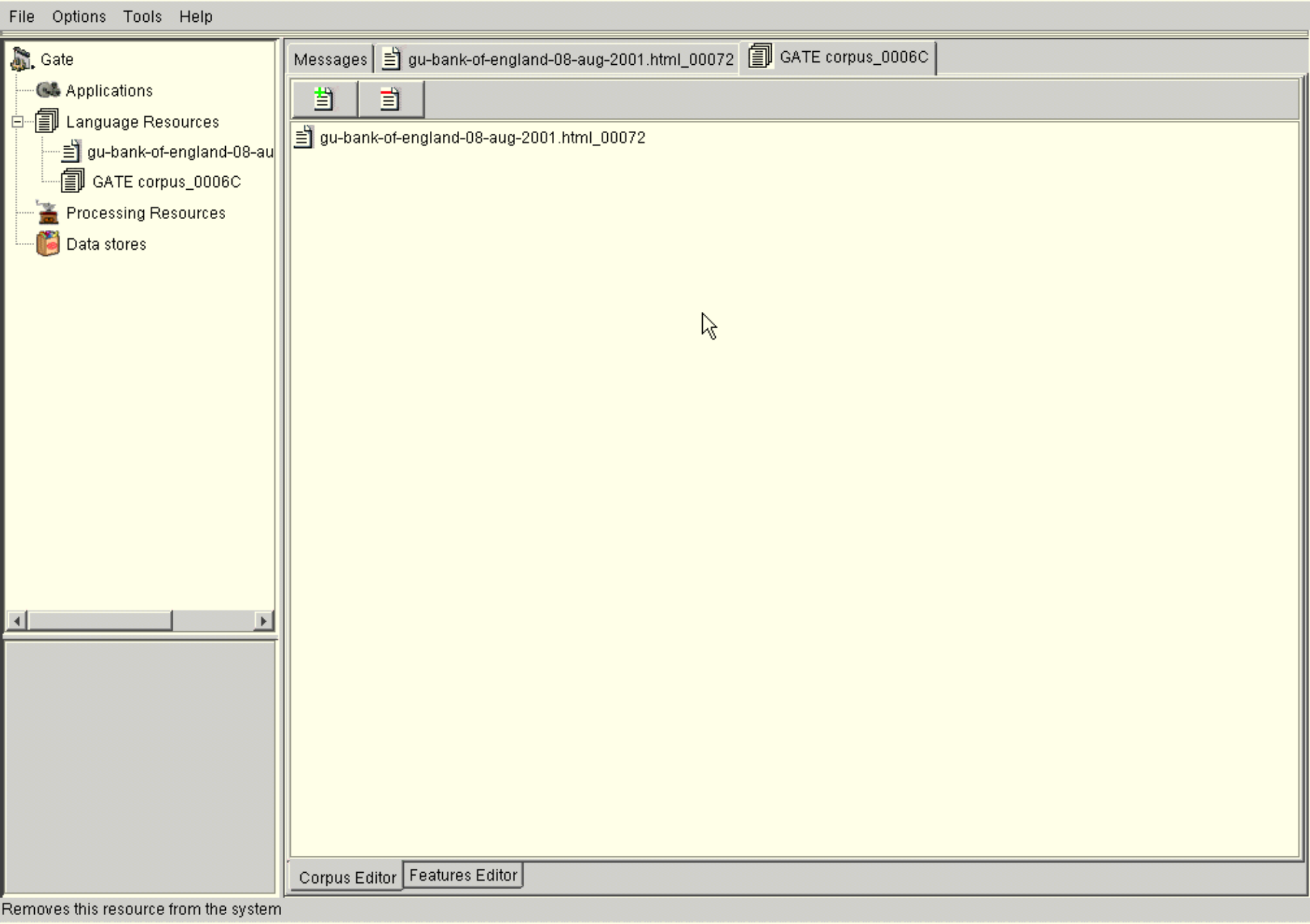

File Options Tools Help
Gate
Applications
Language Resources
example document
Processing Resources
Data stores
Messages
example document
Text
Annotations
Annotation Sets
Print
The Department of Computer Science, University of Sheffield
Latest News Last updated 17th December 2001: DCS is awarded RAE2001 Grade 5
Prospective Students | Current Students | Staff | Visitors
Dept Contact Details | Site Map | Search
University
The Department of Computer Science
Regent Court
211 Portobello Street
Sheffield
S1 4DP
UNITED KINGDOM Tel: +44 (0) 114 22 21800
Fax: +44 (0) 114 22 21810
Email: dept@dcs.shef.ac.uk
© The Department of Computer Science, University of Sheffield 1999
This page is maintained by WebMaster .
Default annotations
Original markups annotations
a
b
body
br
center
div
font
head
hr
html
i
img
link
meta
p
script
table
td
title
tr
Annotations Editor
Features Editor
Removes this resource from the system

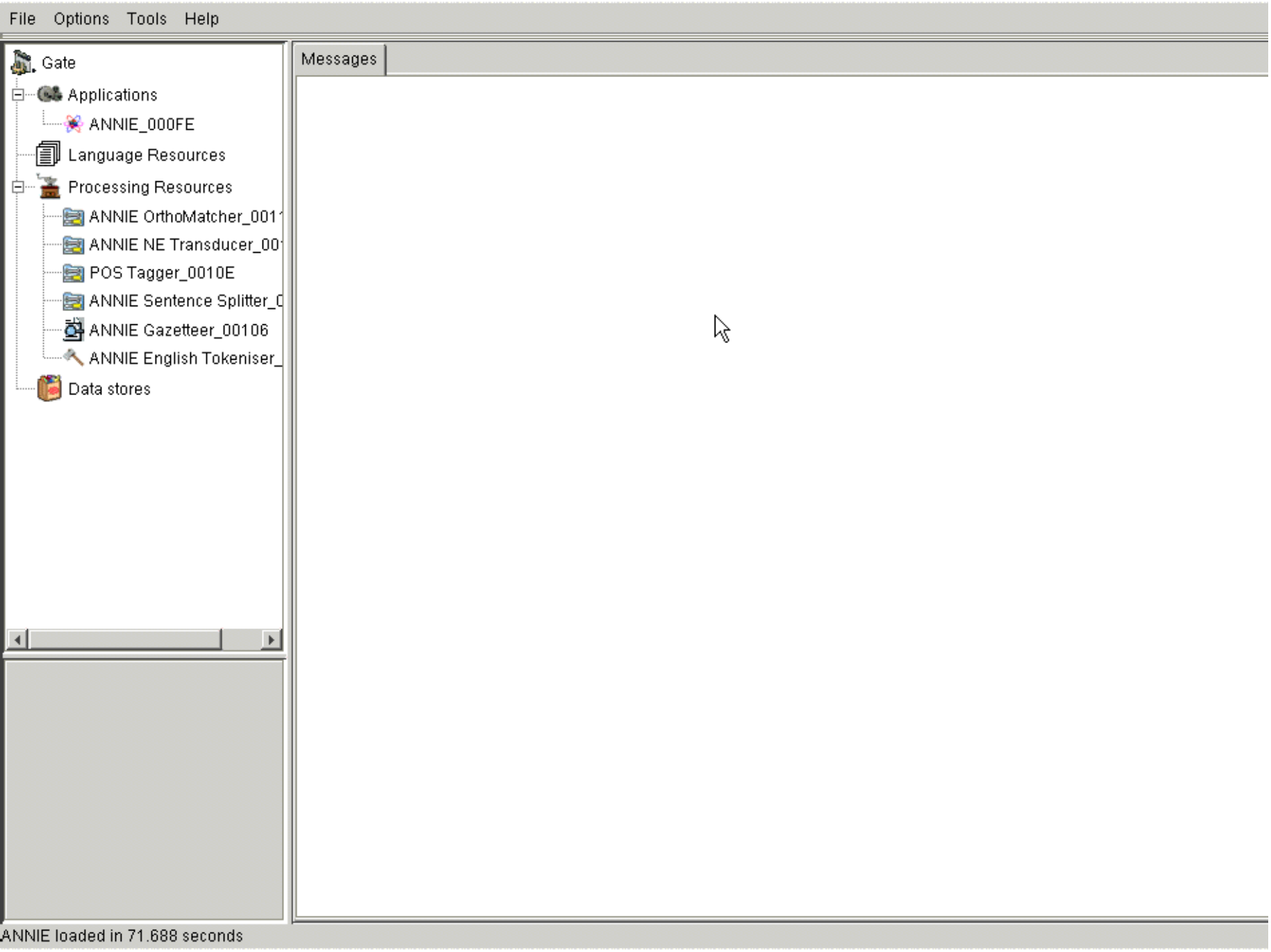

File Options Tools Help
Gate
Applications
Language Resources
news document
Processing Resources
POS Tagger_0005A
ANNIE Sentence Splitter_
ANNIE English Tokeniser_
Data stores
Messages
news document
Text
Annotations
Annotation Sets
Print
9801.175
NEWS STORY
SOURCE: The Washington Post
SECTION: D01
LENGTH: 569
DATE: January 22, 1998
HEADLINE: Roe v. Wade Foes 'Still Going Strong'; Anti rtion Activists Will March to Mark 25th Anniversary of Decision
BODY_LEN: 553
Susan Valentine, a 38-year-old mother of five from Annandale, doesn't care whether 20,000 or 100,000 antiabortion demonstrators turn out for today's "March for Life," the protest held each year in Washington since the U.S. Supreme Court legalized abortion 25 years ago.
What matters, said Valentine, who has been active in the antiabortion movement since she was 13, is that a quarter of a century after the court's Roe v. Wade decision, "we are still going strong, still drawing thousands of people to marches, still making this an issue for candidates. After all this time, it shows the strength of people's convictions."
Organizers of the protest, which will begin with a noon rally at the Ellipse, followed by a march up Constitution Avenue to the Supreme Court building, declined yesterday to predict how many supporters will attend.
But they said the 25th anniversary will draw a larger turnout than the 50,000 to 75,000 who have participated in recent years.
Default annotations
Original markups annotations
DOC
DOCNO
DOCTYPE
HEADER
TEXT
Annotations Editor
Features Editor
loaded in 8.094 seconds

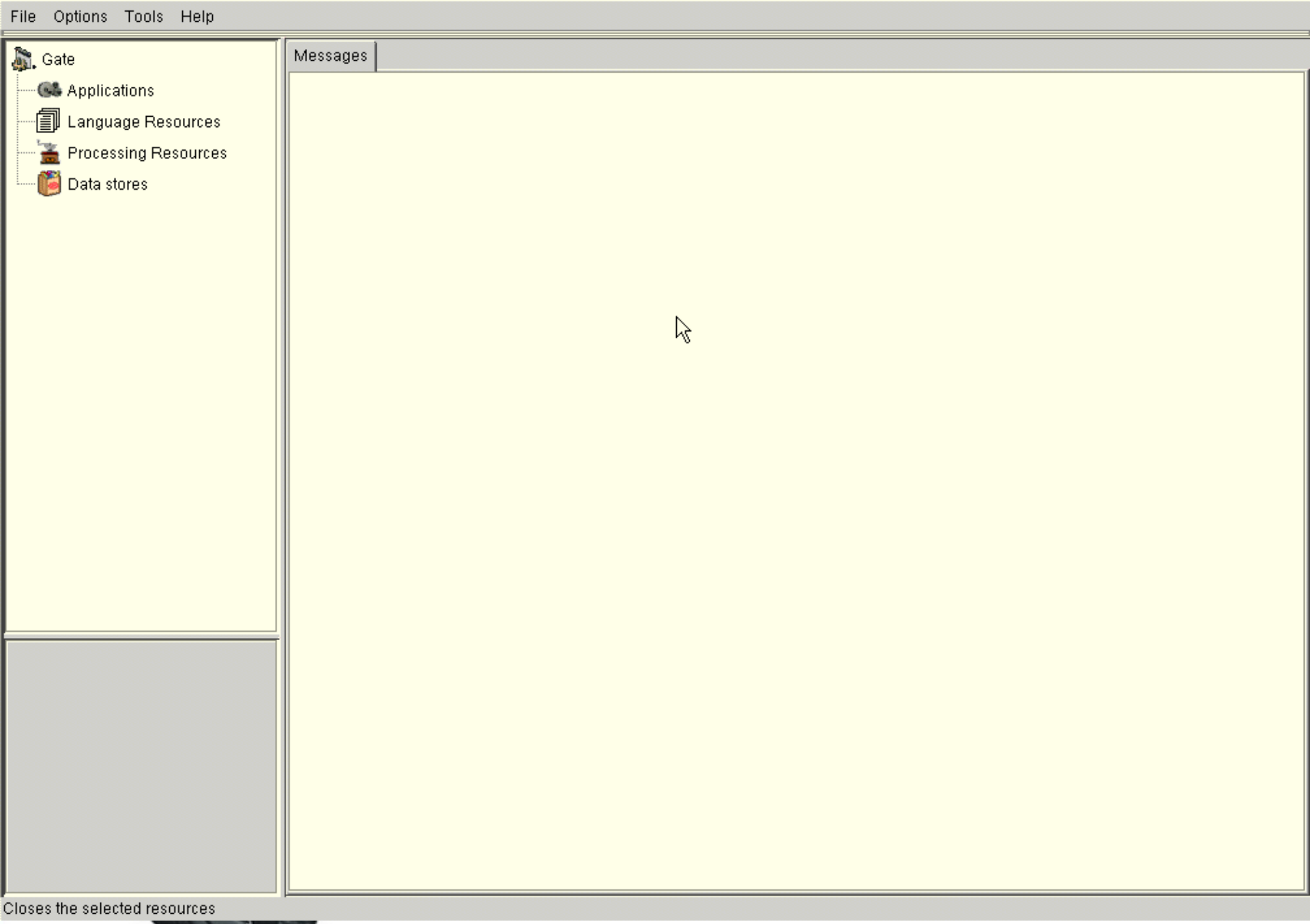

File
Options
Tools
Help
Gate
Applications
Language Resources
news document
Processing Resources
Data stores
Messages

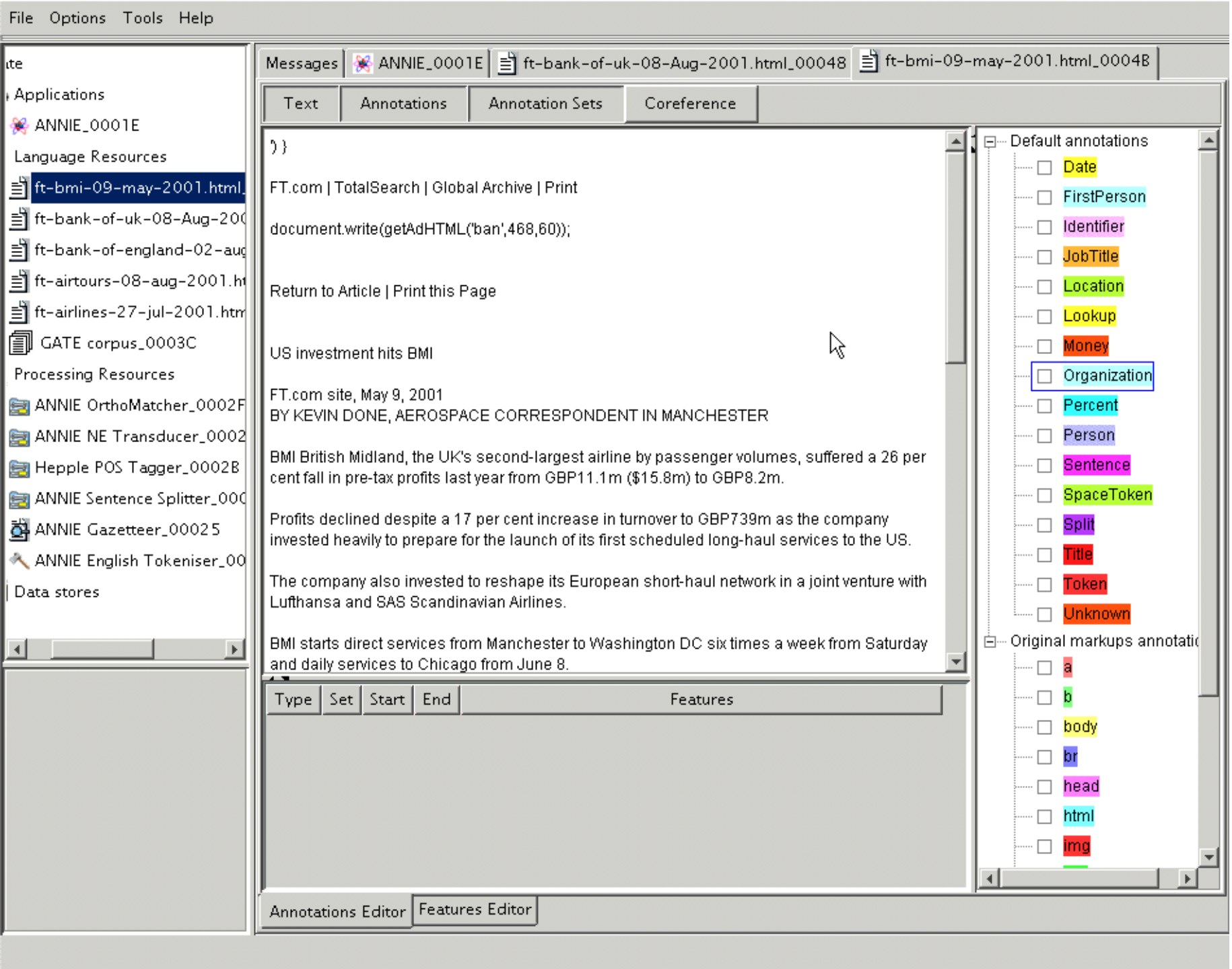

File
Options
Tools
Help
Gate
Applications
Language Resources
GATE corpus_0006C
Processing Resources
Data stores
Messages

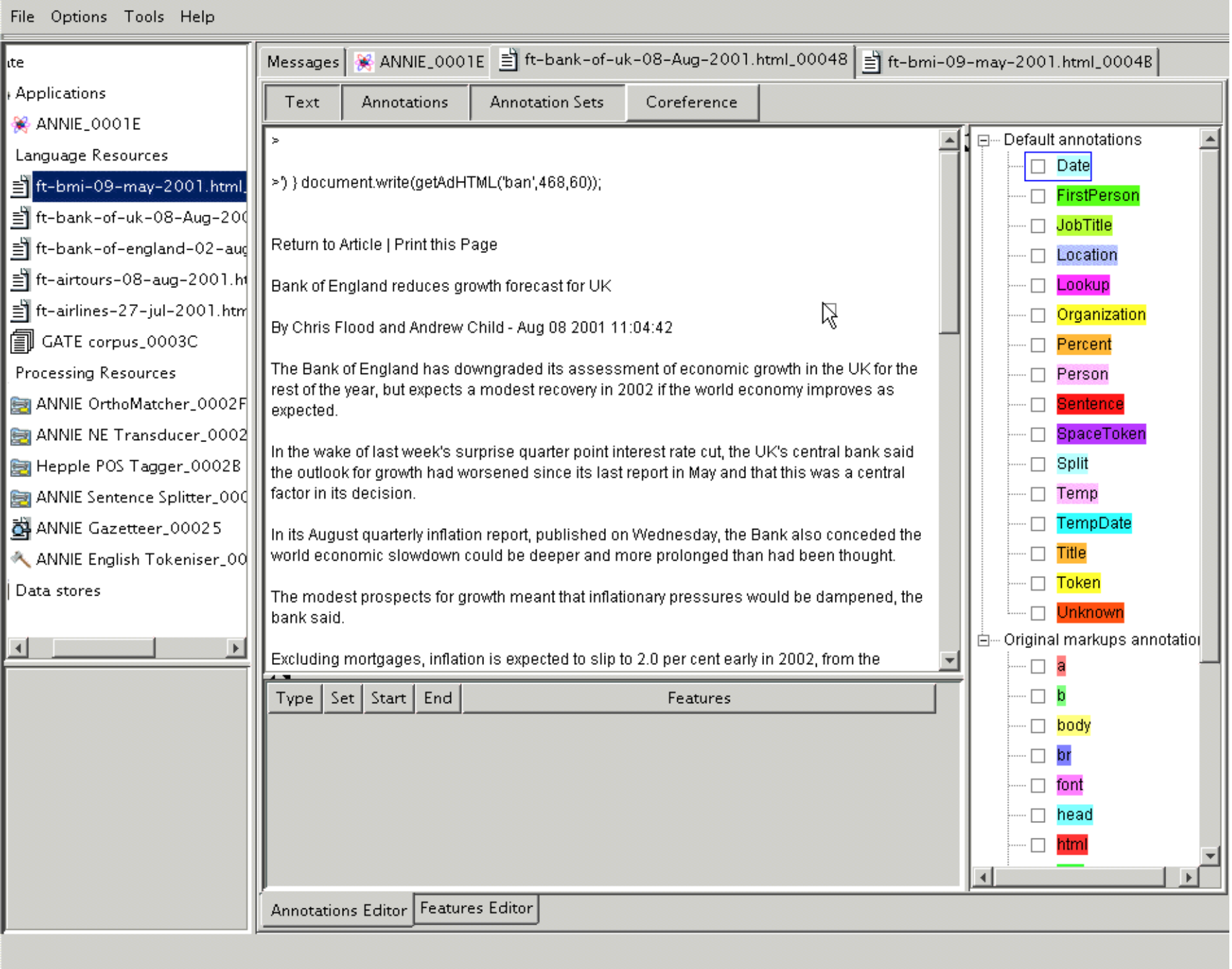

File
Options
Tools
Help
Gate
Applications
Language Resources
gu-bank-of-england-08-au
GATE corpus_0006C
Processing Resources
POS Tagger_00088
ANNIE Sentence Splitter_0
ANNIE English Tokeniser_
Data stores
Messages
gu-bank-of-england-08-aug-2001.html_00072
GATE corpus_0006C
gu-bank-of-england-08-aug-2001.html_00072
Corpus Editor
Features Editor

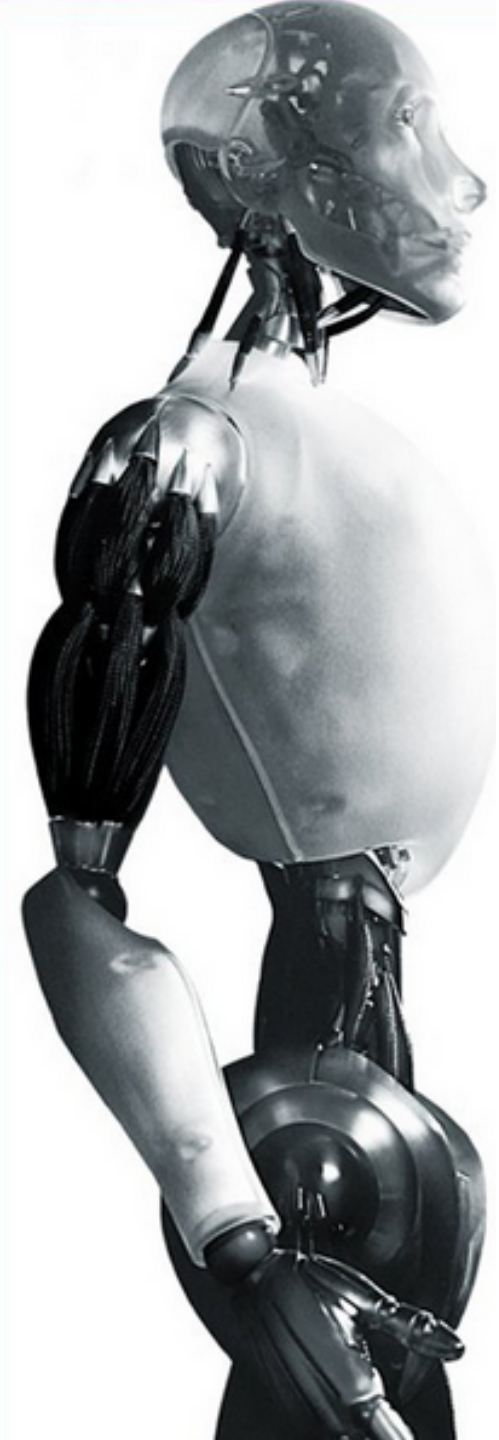

File
Options
Tools
Help
Gate
Applications
Language Resources
gu-bank-of-england-08-au
GATE corpus_0006C
Processing Resources
Data stores
Messages
gu-bank-of-england-08-aug-2001.html_00072
GATE corpus_0006C
gu-bank-of-england-08-aug-2001.html_00072
Corpus Editor
Features Editor
Removes this resource from the system


File
Options
Tools
Help
Gate
Applications
ANNIE_000FE
Language Resources
Processing Resources
ANNIE OrthoMatcher_0011
ANNIE NE Transducer_00
POS Tagger_0010E
ANNIE Sentence Splitter_0
ANNIE Gazetteer_00106
ANNIE English Tokeniser_
Data stores
Messages
ANNIE loaded in 71.688 seconds


File
Options
Tools
Help
Gate
Applications
Language Resources
Processing Resources
Data stores
Messages
Closes the selected resources

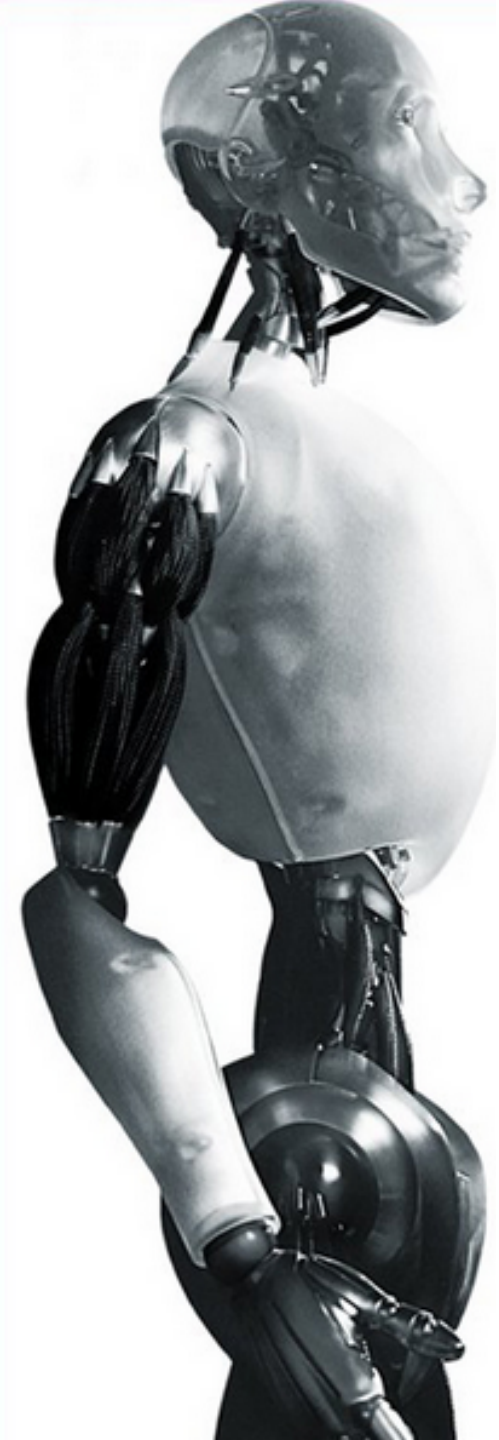

File Options Tools Help
Applications
ANNIE_0001E
Language Resources
ft-bmi-09-may-2001.html
ft-bank-of-uk-08-Aug-200
ft-bank-of-england-02-aug
ft-airtours-08-aug-2001.ht
ft-airlines-27-jul-2001.htm
GATE corpus_0003C
Processing Resources
ANNIE OrthoMatcher_0002F
ANNIE NE Transducer_0002
Hepple POS Tagger_0002B
ANNIE Sentence Splitter_000
ANNIE Gazetteer_00025
ANNIE English Tokeniser_00
Data stores
Messages
ANNIE_0001E
ft-bank-of-uk-08-Aug-2001.html_00048
ft-bmi-09-may-2001.html_0004B
Text
Annotations
Annotation Sets
Coreference
') }
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.
Type
Set
Start
End
Features
Default annotations
Date
FirstPerson
Identifier
JobTitle
Location
Lookup
Money
Organization
Percent
Person
Sentence
SpaceToken
Split
Title
Token
Unknown
Original markups annotati
a
b
body
br
head
html
img
Annotations Editor
Features Editor


File Options Tools Help
Applications
ANNIE_0001E
Language Resources
ft-bmi-09-may-2001.html
ft-bank-of-uk-08-Aug-200
ft-bank-of-england-02-aug
ft-airtours-08-aug-2001.ht
ft-airlines-27-jul-2001.htm
GATE corpus_0003C
Processing Resources
ANNIE OrthoMatcher_0002F
ANNIE NE Transducer_0002
Hepple POS Tagger_0002B
ANNIE Sentence Splitter_000
ANNIE Gazetteer_00025
ANNIE English Tokeniser_00
Data stores
Messages
ANNIE_0001E
ft-bank-of-uk-08-Aug-2001.html_00048
ft-bmi-09-may-2001.html_0004B
Text
Annotations
Annotation Sets
Coreference
>
>') } document.write(getAdHTML('ban',468,60));
Return to Article | Print this Page
Bank of England reduces growth forecast for UK
By Chris Flood and Andrew Child - Aug 08 2001 11:04:42
The Bank of England has downgraded its assessment of economic growth in the UK for the rest of the year, but expects a modest recovery in 2002 if the world economy improves as expected.
In the wake of last week's surprise quarter point interest rate cut, the UK's central bank said the outlook for growth had worsened since its last report in May and that this was a central factor in its decision.
In its August quarterly inflation report, published on Wednesday, the Bank also conceded the world economic slowdown could be deeper and more prolonged than had been thought.
The modest prospects for growth meant that inflationary pressures would be dampened, the bank said.
Excluding mortgages, inflation is expected to slip to 2.0 per cent early in 2002, from the
Type
Set
Start
End
Features
Default annotations
Date
FirstPerson
JobTitle
Location
Lookup
Organization
Percent
Person
Sentence
SpaceToken
Split
Temp
TempDate
Title
Token
Unknown
Original markups annotation
a
b
body
br
font
head
html
Annotations Editor
Features Editor



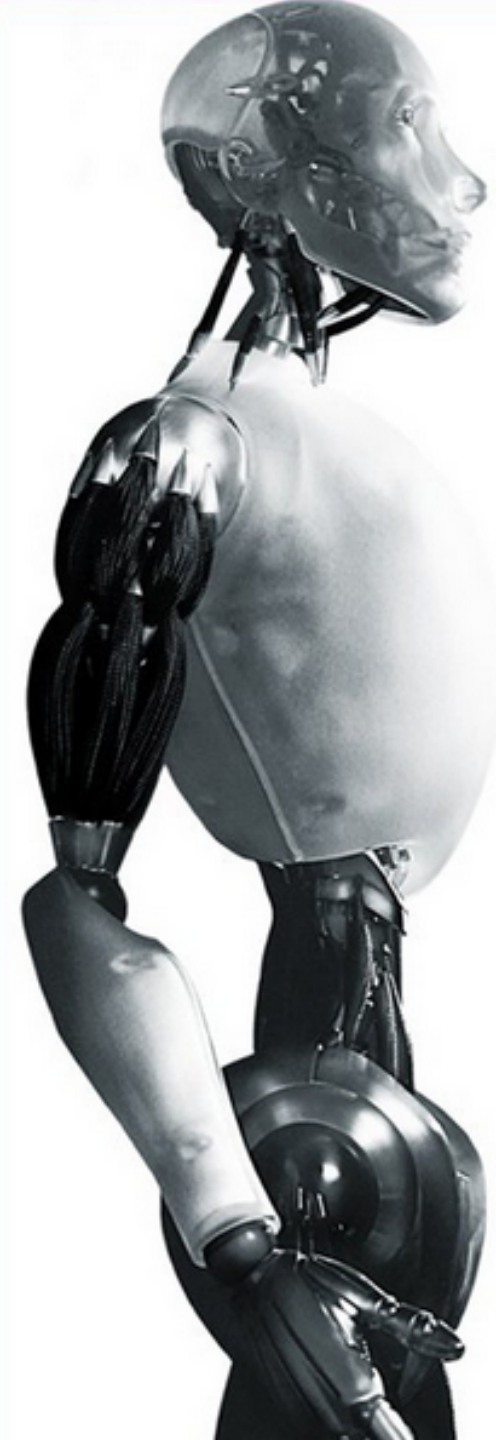
The Gazetteer

- Gazetteers are plaintext files containing lists of names (e.g rivers, cities, people, ...)
- Information used by JAPErules
- Each gazetteer set has an index file listing all the lists, plus features of each list (majorType, minorType and language)
- Lists can be modified either internally using Gate, or externally in your favorite editor
- Generates Lookup results of the given kind



The Lists

- Set of lists compiled into Finite State Machines
- 60k entries in 80 types, inc.: organization; artifact; location; amount_unit; manufacturer; transport_means; company_designator; currency_unit; date; government_designator; ...
- Each list has attributes MajorType and MinorType and Language): city.lst: location: city: english currency_prefix.lst: currency_unit: pre_amount currency_unit.lst: currency_unit: post_amount
- List entries may be entities or parts of entities, or they may contain contextual information (e.g. job titles often indicate people)



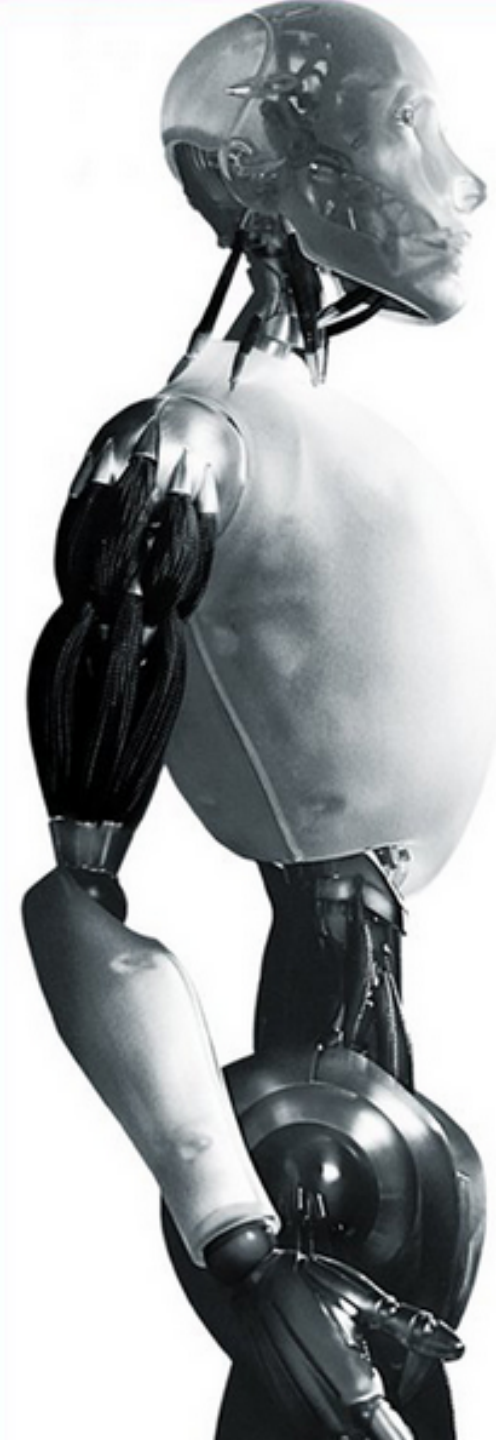
Gazetteer - Exercise

- Download the introduction of this article …

  http://mt.wikipedia.org/wiki/Malta

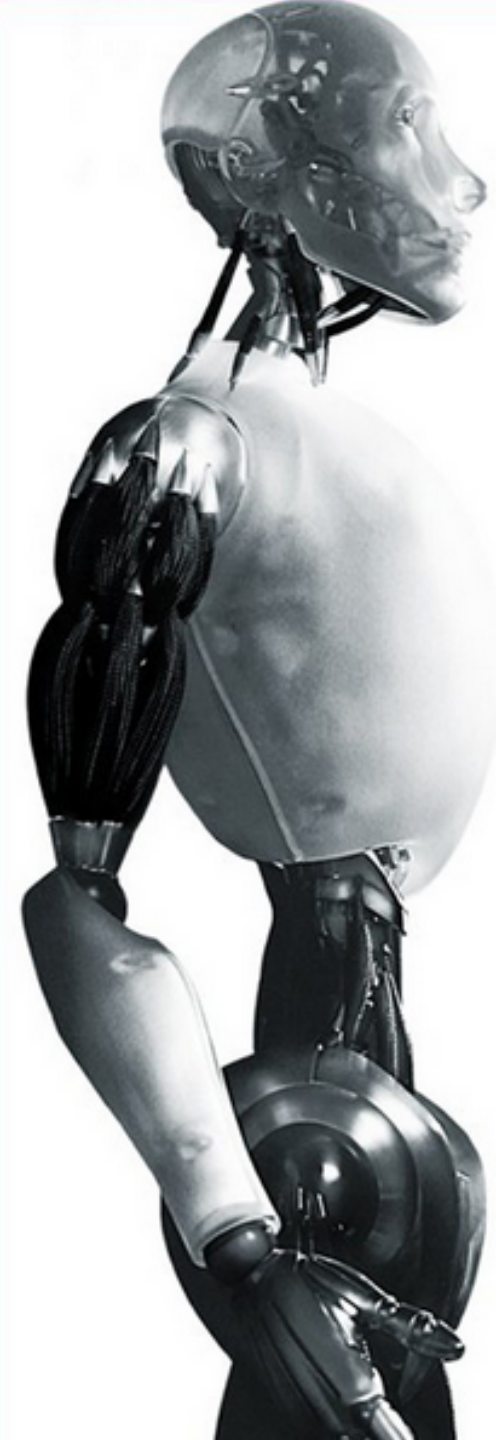
- Add to the Gazetteer elements which have been missed such as names, locations, etc.

- Add the Pronominal Co-reference to the pipeline … what is its effect?



JAPE grammars

- A semantic tagger consists of a set of rule-based JAPE grammars run sequentially
- JAPE is a pattern-matching language
- The LHS of each rule contains patterns to be matched
- The RHS contains details of annotations (and optionally features) to be created
- More complex rules can also be created



Input specifications

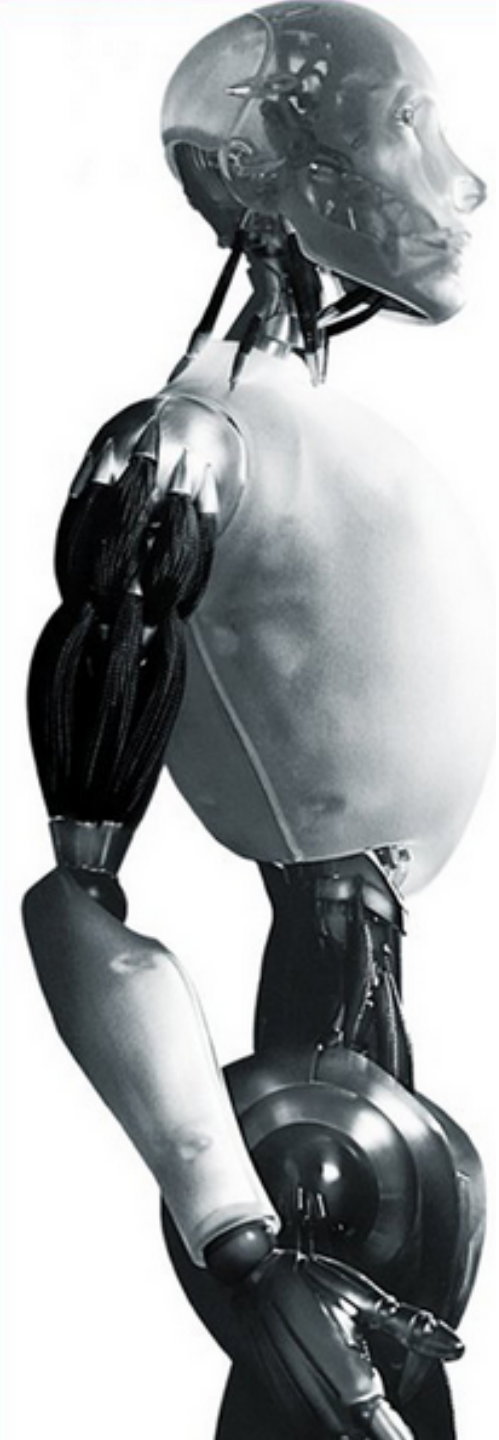
- The head of each grammar phase needs to contain certain information
  - Phase name
  - Inputs
  - Matching style

  e.g.

Phase: location

Input: Token Lookup Number

Control: appelt



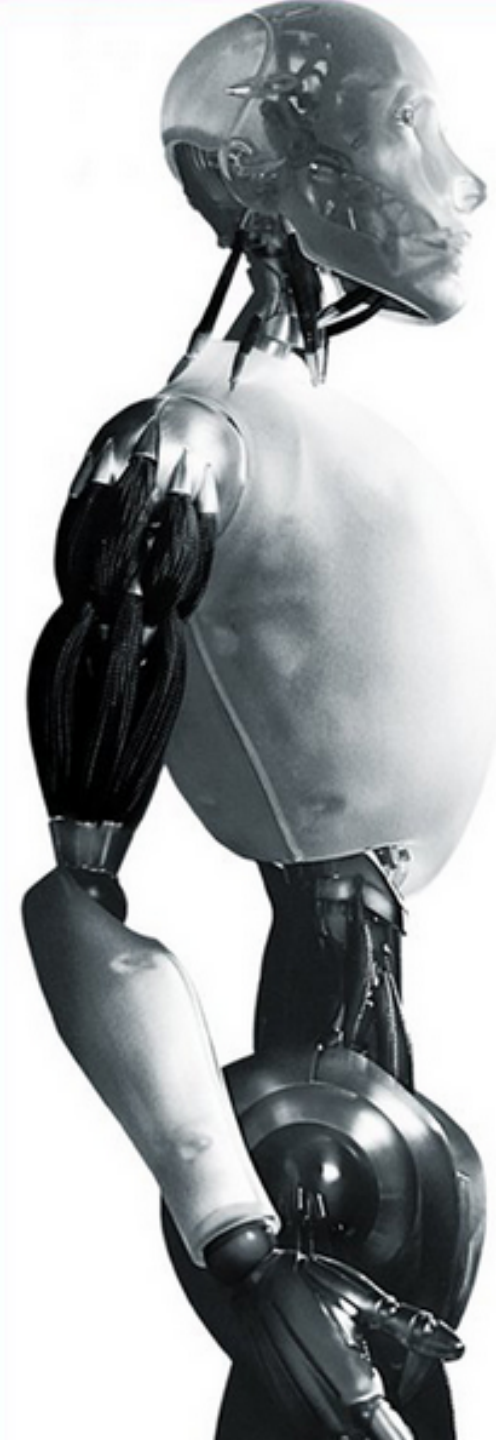
Matching algorithms and Rule Priority

- 3 styles of matching:
  - Brill (fire every rule that applies)
  - First (shortest rule fires)
  - Appelt (use of priorities)

- Appelt priority is applied in the following order
  - Starting point of a pattern
  - Longest pattern
  - Explicit priority (default = -1)

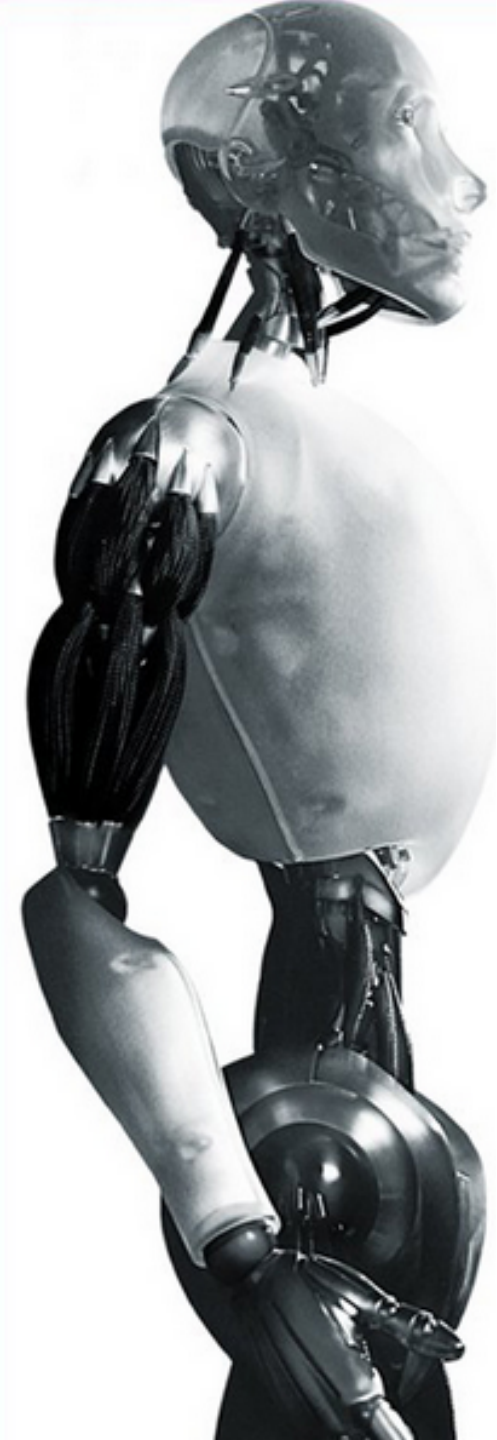
NE Rule in JAPE

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



LHS of the rule

- LHS is expressed in terms of existing annotations, and optionally features and their values
- Any annotation to be used must be included in the input header
- Any annotation not included in the input header will be ignored (e.g. whitespace)
- Each annotation is enclosed in curly braces
- Each pattern to be matched is enclosed in round brackets and has a label attached



Macros

- Macros look like the LHS of a rule but have no label

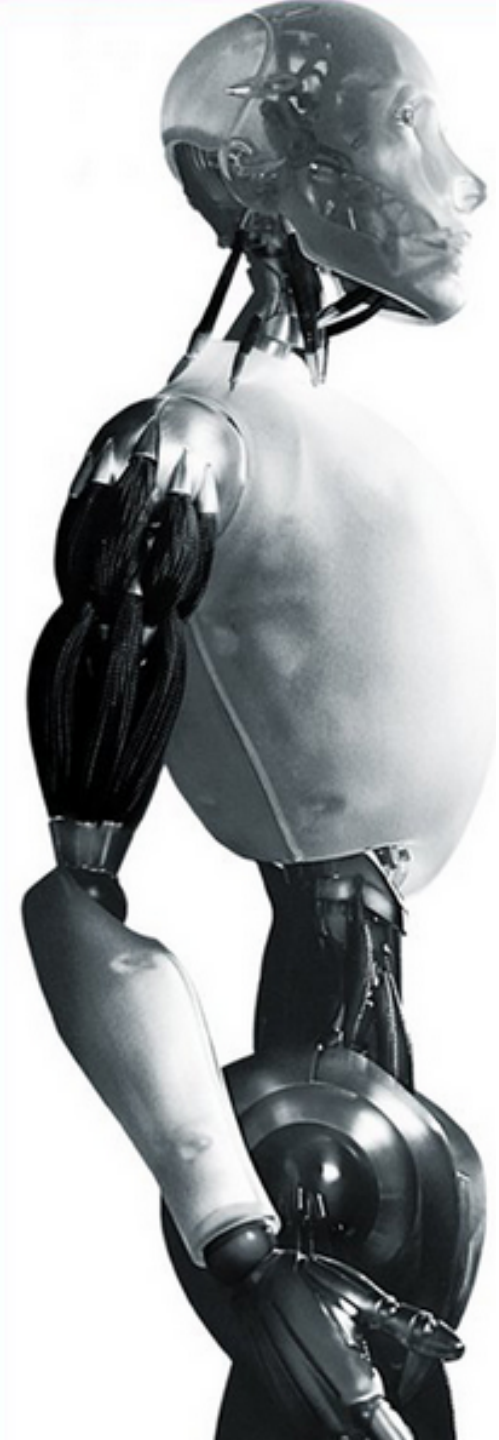
Macro: NUMBER

(({Digit})+)

- They are used in rules by enclosing the macro name in round brackets

((NUMBER)+):match

- Conventional to name macros in uppercase letters
- Macros hold across an entire set of grammar phases



Contextual information

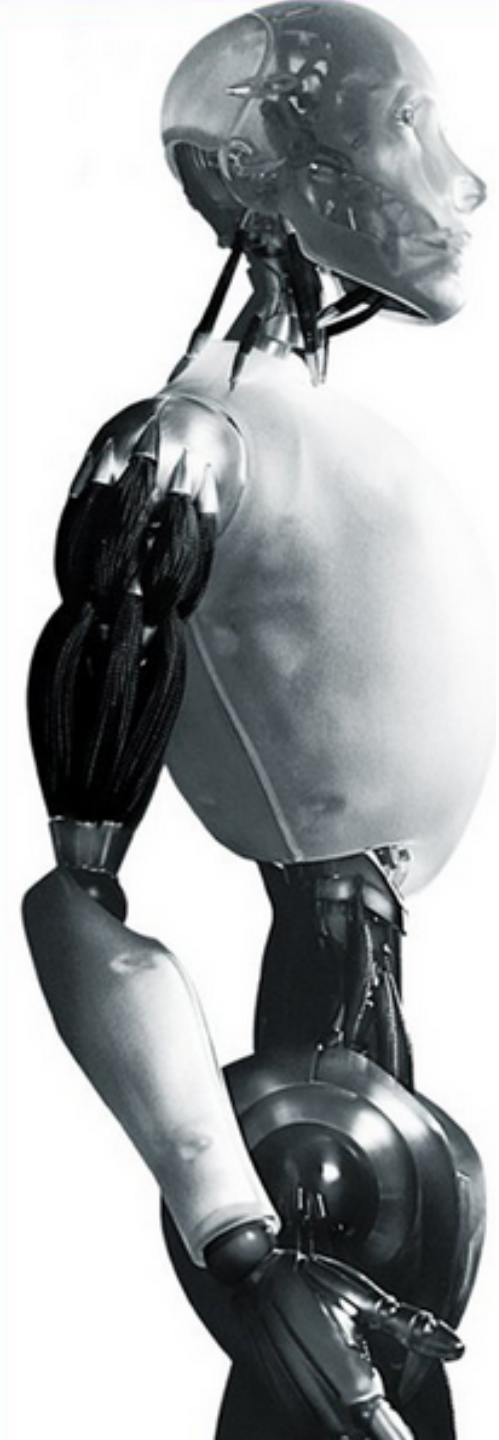
- Contextual information can be specified in the same way, but has no label
- Contextual information will be consumed by the rule

({Annotation1})

({Annotation2}):match

({Annotation3})

→

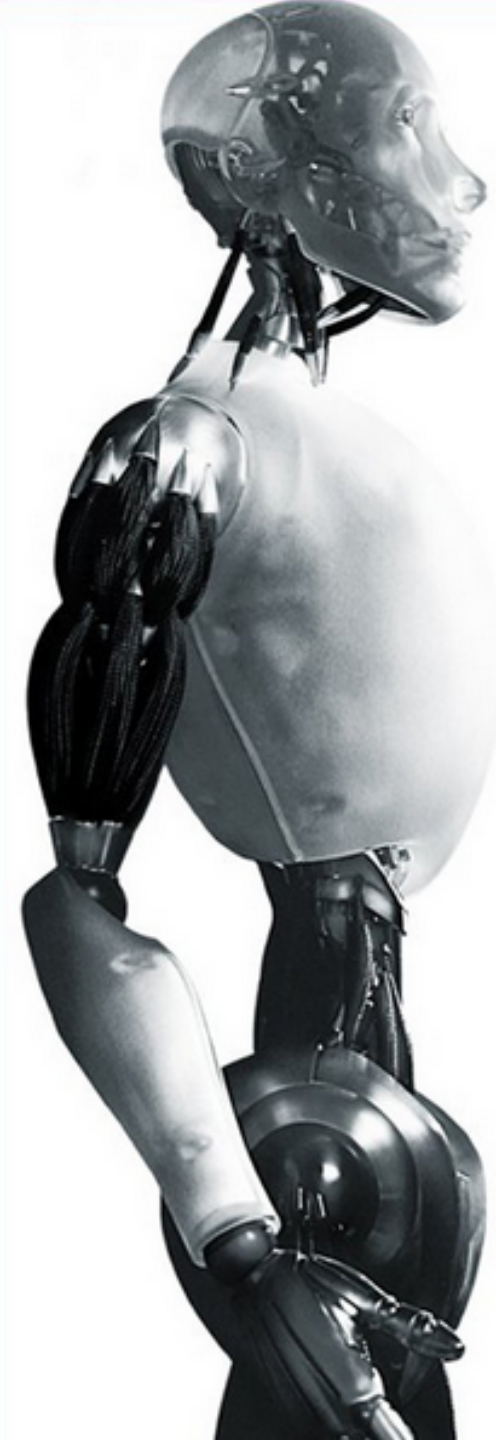


RHS of the rule

- LHS and RHS are separated by →
- Label matches that on the LHS
- Annotation to be created follows the label

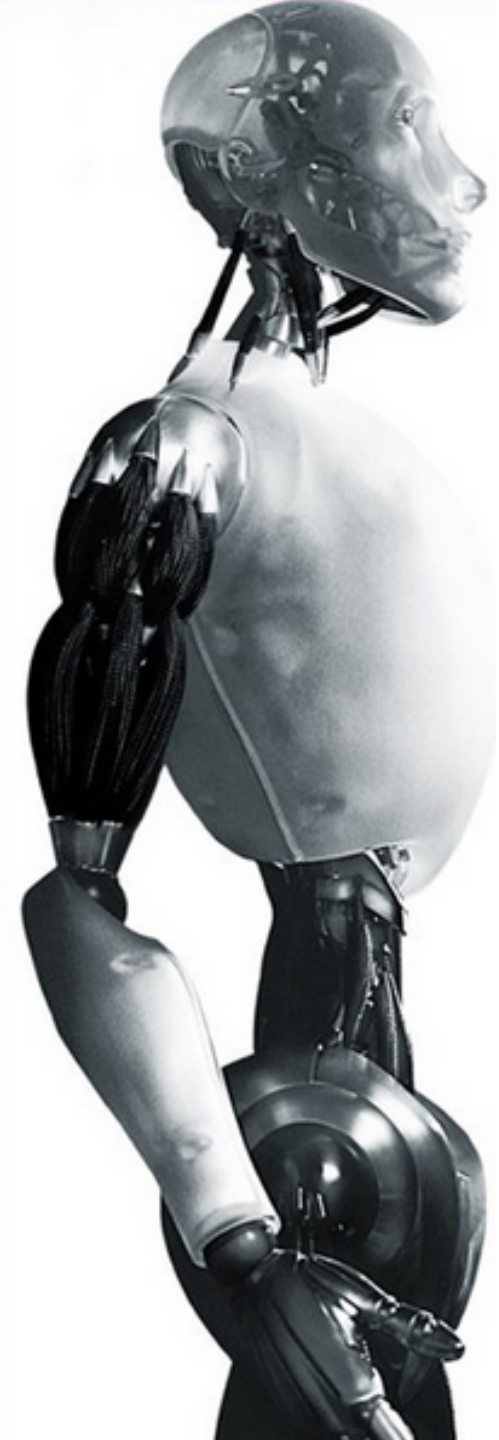
(Annotation1):match

→ :match.NE = {feature1 = value1, feature2 = value2}



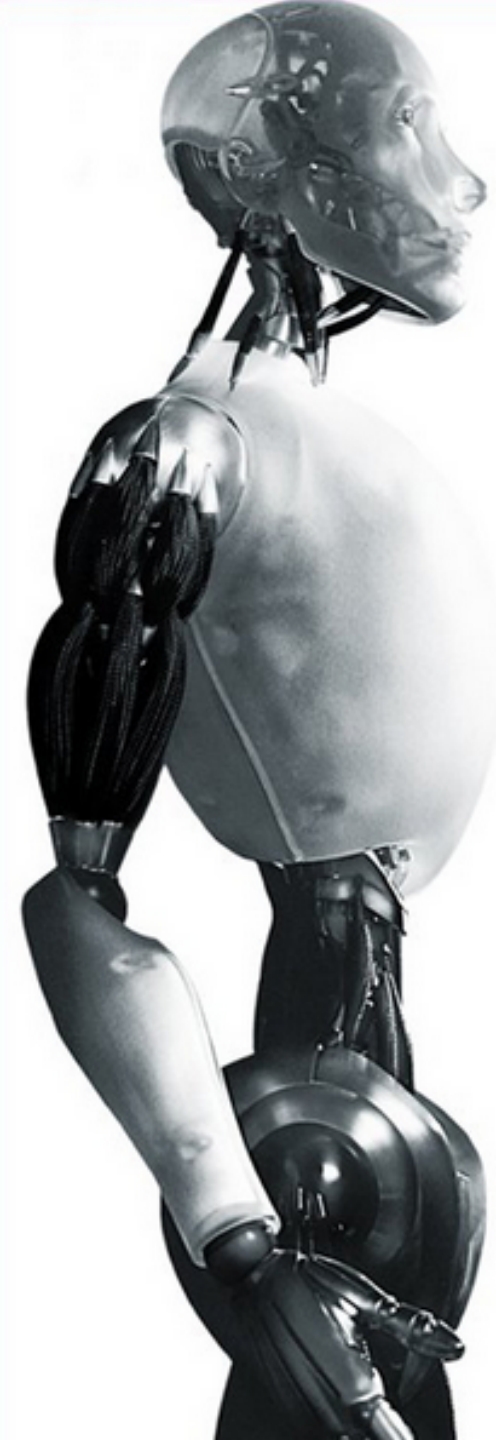
More complex JAPE rules

- Any Java code can be used on the RHS of a rule
- This is useful for e.g. feature percolation, ontology population, accessing information not readily available, comparing feature values, deleting existing annotations etc.
- There are examples of these in the user guide and in the ANNIE NE grammars
- Most JAPE rules end up being complex!



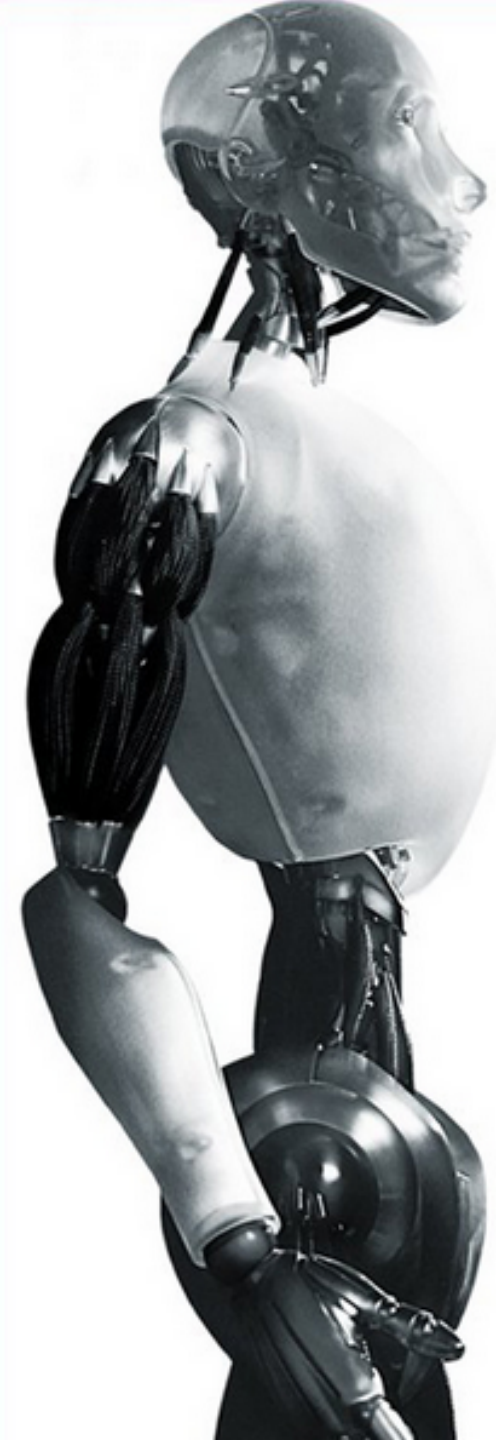
Using JAPE for other tasks

- JAPE grammars are not just useful for NE annotation
- They can be a quick and easy way of performing any kind of task where patterns can be easily recognised and a finite-state approach is possible, e.g. transforming one style of markup into another, deriving features for the learning algorithms



Exercise

- Create a JAPE rule which matches the text “University of Malta” or any other University and annotates it with a tag kind = university



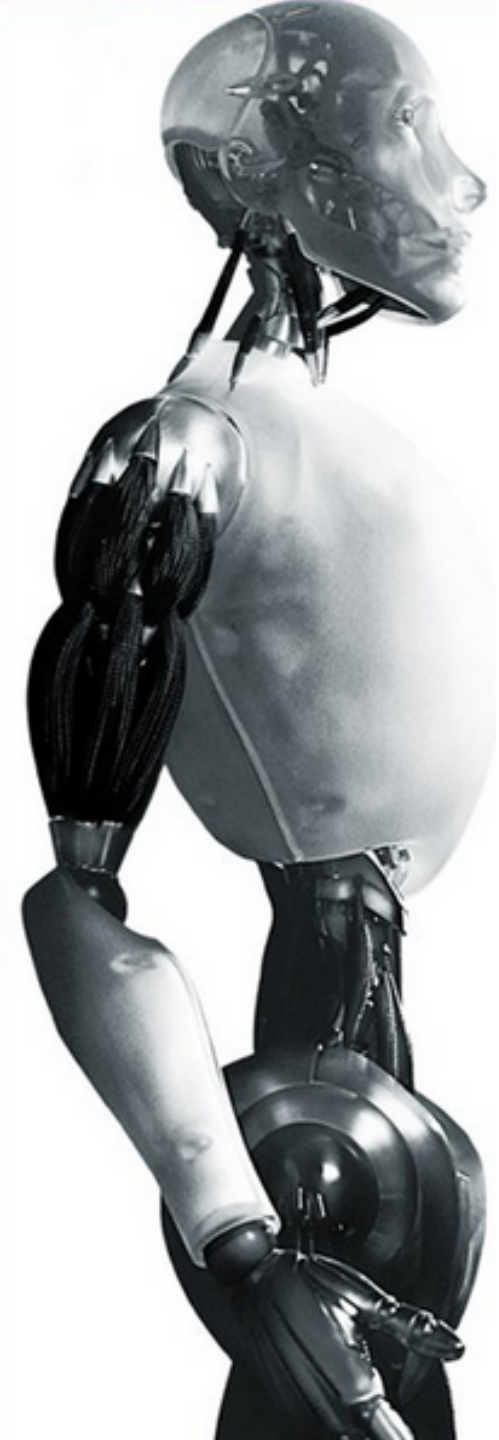
Answer

Rule: University1

({Token.string == "University"}
{Token.string == "of"}
{Lookup.minorType ==
city}):orgName

-->

:orgName.Organisation = {kind =
"university", rule = "University1"}



Questions?