Introduction to the GATE language analysis system

Wim Peters
Natural Language Processing Group
Department of Computer Science
University of Sheffield

Historical Text Mining Workshop,
Lancaster July 20-21, 2006

Some scholarly requirements for computational historical linguistics

- Annotation of texts
  manual/automatic ; stand alone/collaborative

- Analysis types include:
  - Tokenization
  - Part of speech tagging
  - Parsing
  - Morphological analysis
  - Ngram extraction/comparison
  - Identification of cognates between languages
  - Semantic comparison of cognates

Requirements continued

- Visualisation: view textual resources and results of annotation processes
  - Lemma/wordform list; frequency counts
  - Bigrams/trigrams etc.
  - Concordancing
  - Evaluation of results
  - Export results in XML
  - Search text and annotations

GATE

- GATE (Generalised Architecture for Text Engineering)

- is a framework for language processing
- under constant development on the basis of EU and national funding

- includes language processing tools, e.g. pos taggers, parsers for various languages
- tools for visualising and manipulating ontologies
- ontology-based information extraction tools
- evaluation tools

- is freely available (www.gate.ac.uk)

GATE Users

- **American National Corpus** project
- **Perseus Digital Library** project, Tufts University, US
- **Longman Pearson** publishing, UK
- **Merck KgAa**, Germany
- **Canon Europe**, UK
- **a large number of UK, US and EU Universities**
- **UK and EU projects** include
  - **EMILLE**: S. Asian languages corpus
  - **ETCSL** ancient Sumerian corpus
  - **Old Bailey**: 17th century court reports
  - **ACE / TIDES**: semantic annotation for Arabic and Chinese

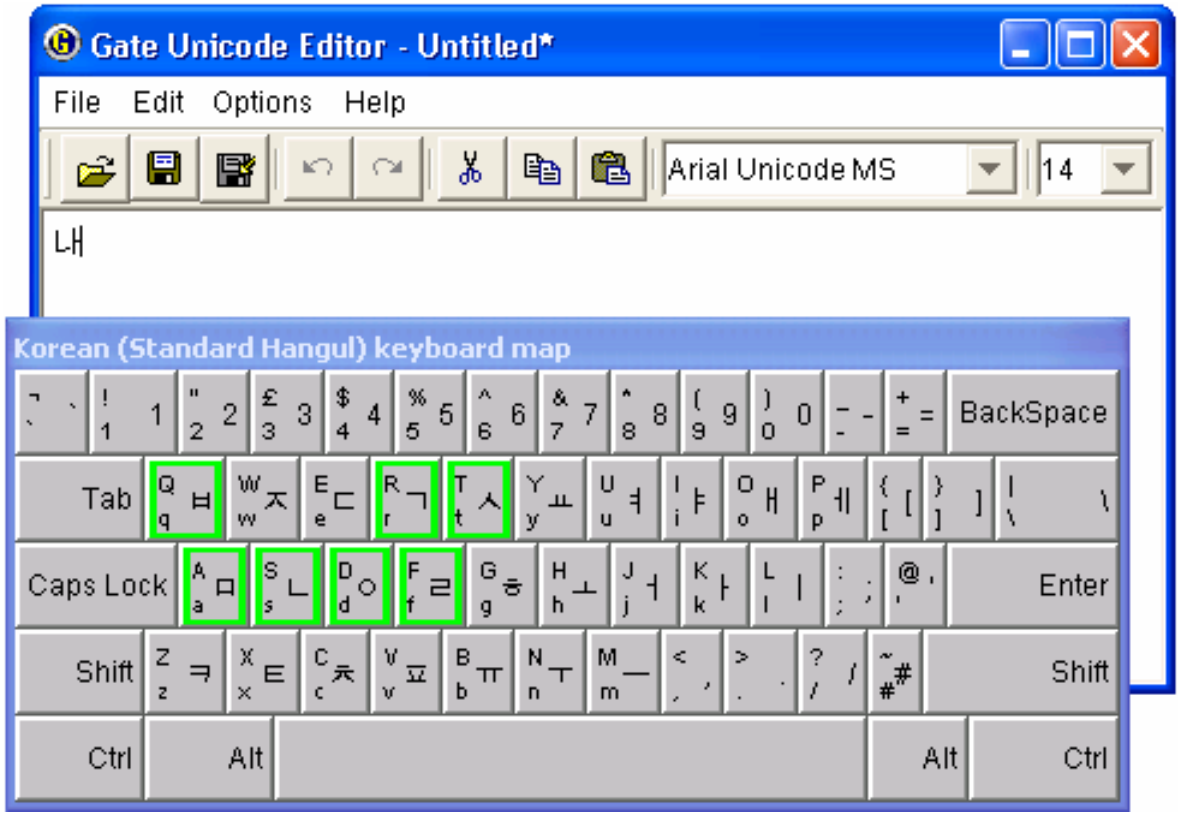
GATE modules I: Language Resources

GATE LRs are documents, ontologies, corpora, lexicons.

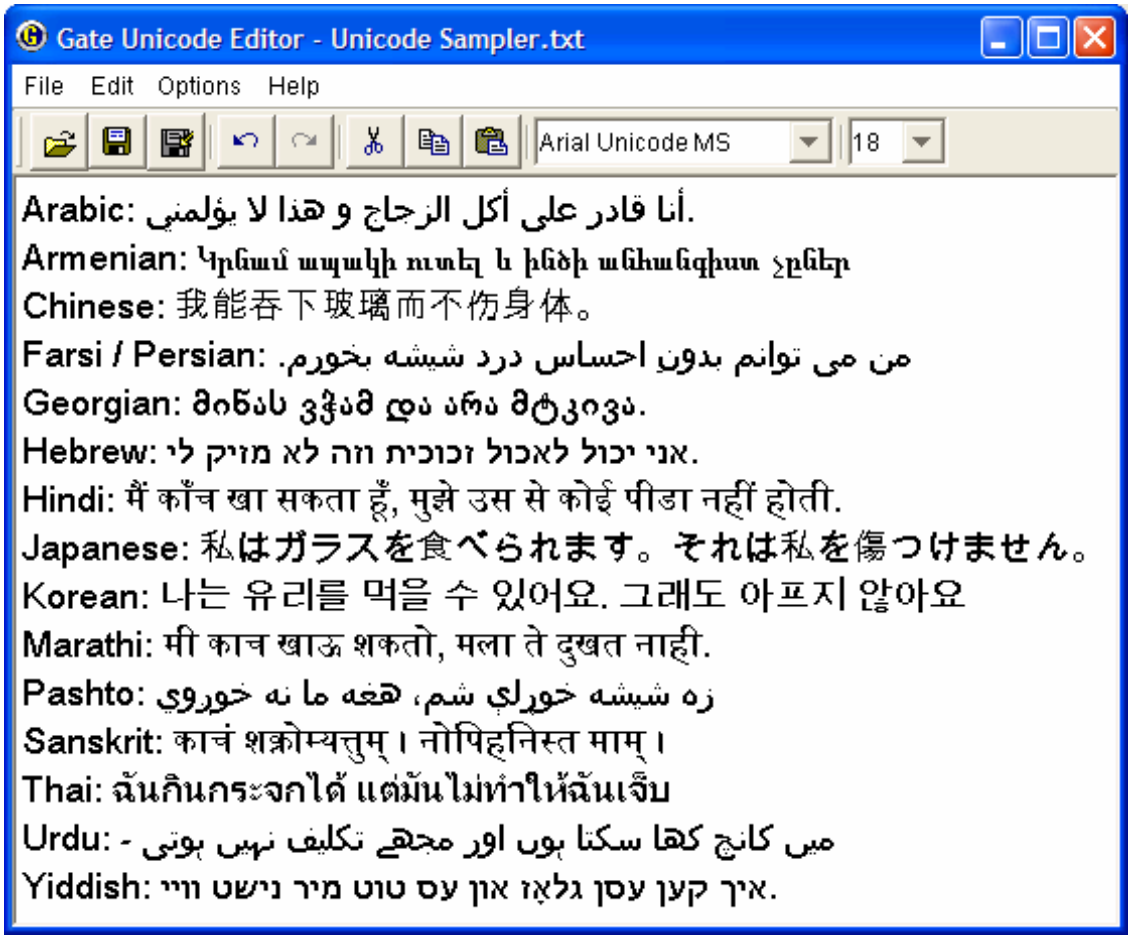
Documents / corpora:
• GATE documents loaded from local files or the web;
• Diverse document formats: text, html, XML, email, RTF, SGML.

Multilinguality:
•Unicode compliant
• virtual keyboards for language input

Editing Multilingual Data



Displaying Multilingual Data

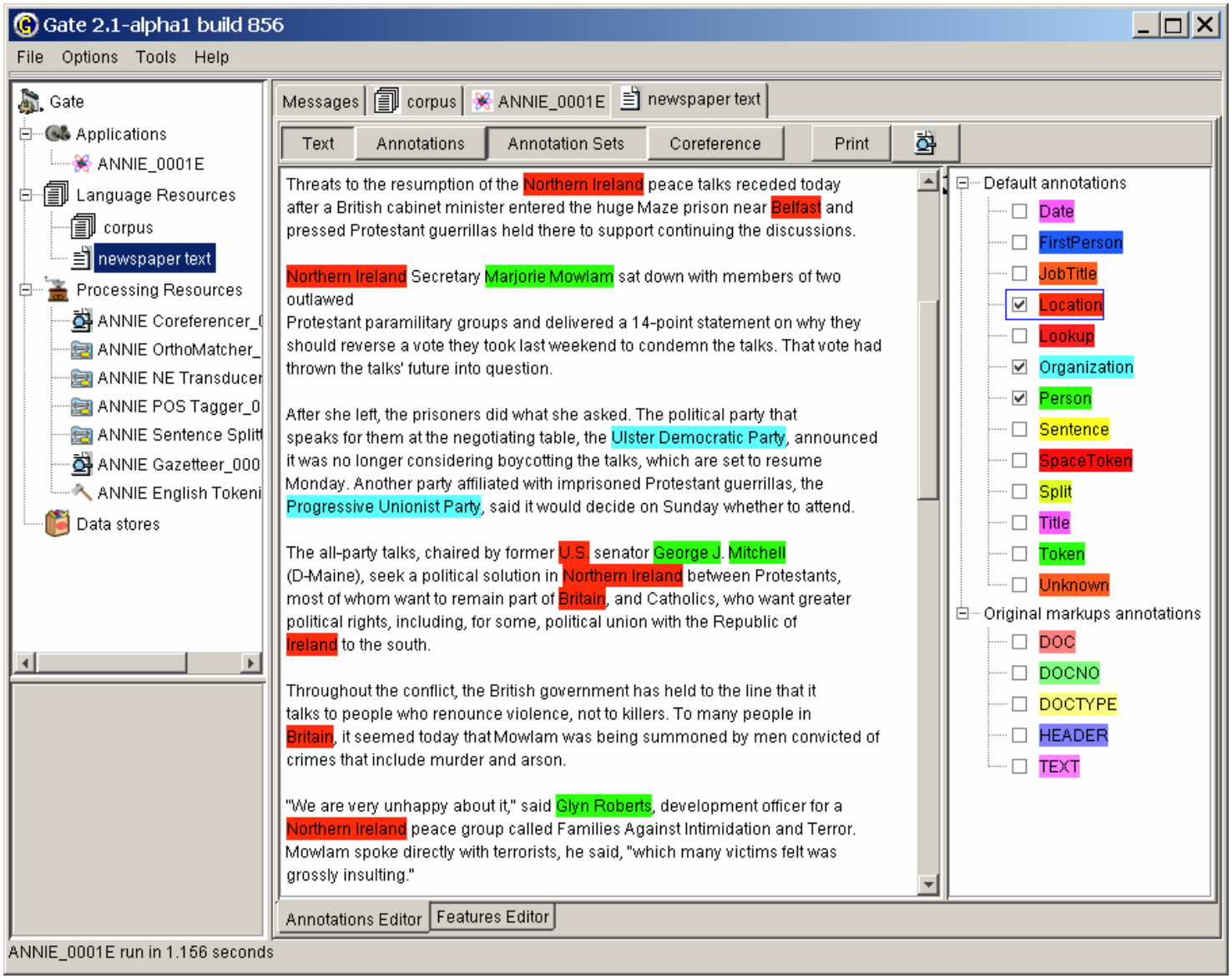


GATE modules II: Processing Resources

- Programs that run over texts and add annotations.
- All PRs can handle Unicode data by default.
- Quite a few freely available with GATE, e.g.
    - Tokeniser
    - Sentence splitter
    - Part of speech tagger
    - Morphological analyser
    - Parser
    - Semantic annotator
- New programs can be:
    - created within GATE
    - Integrated into Gate by means of wrappers
- Users are constantly providing new modules.
- The results of running PRs over texts can be exported in XML.

- Quick demo?

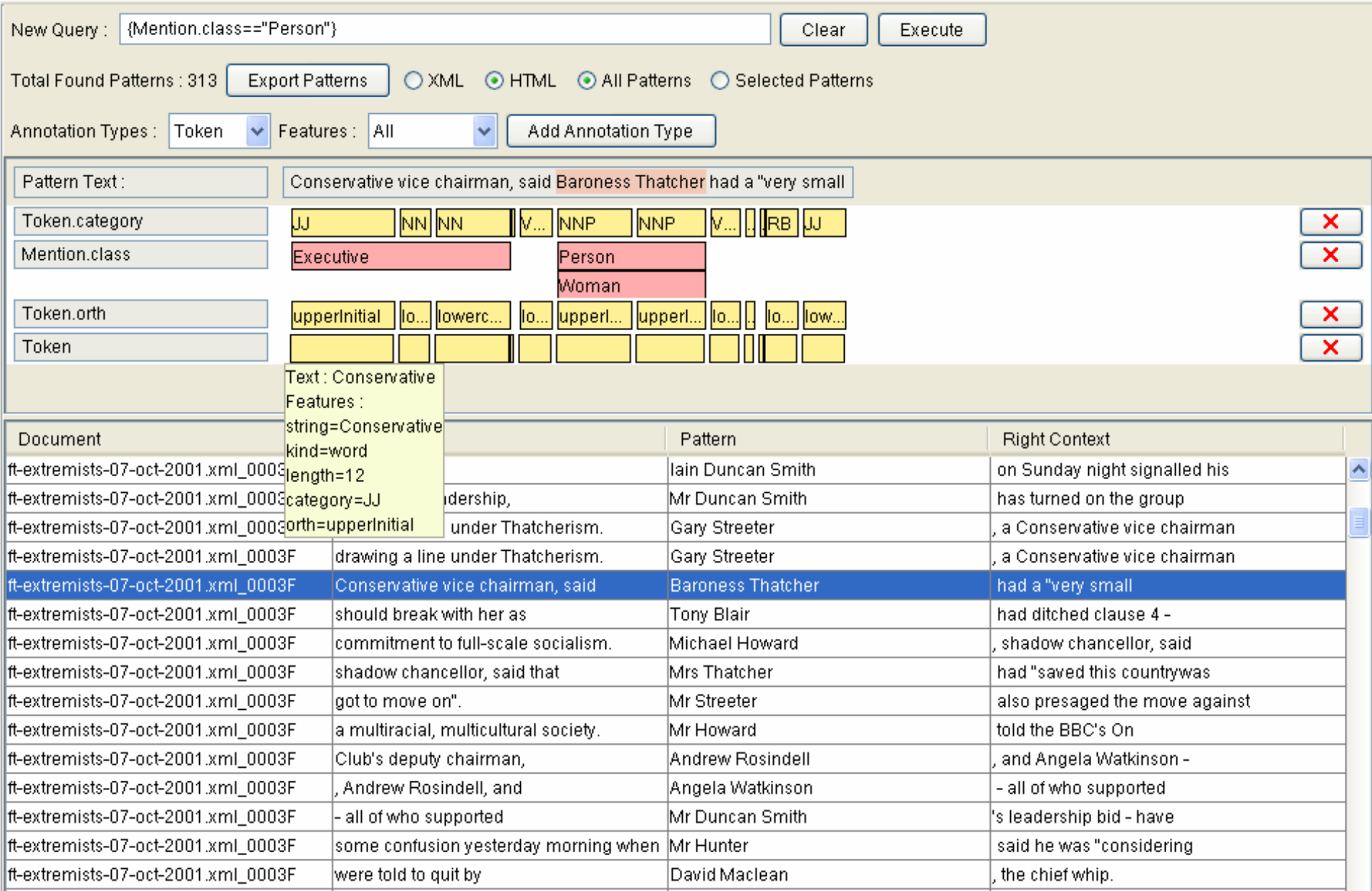
GATE


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

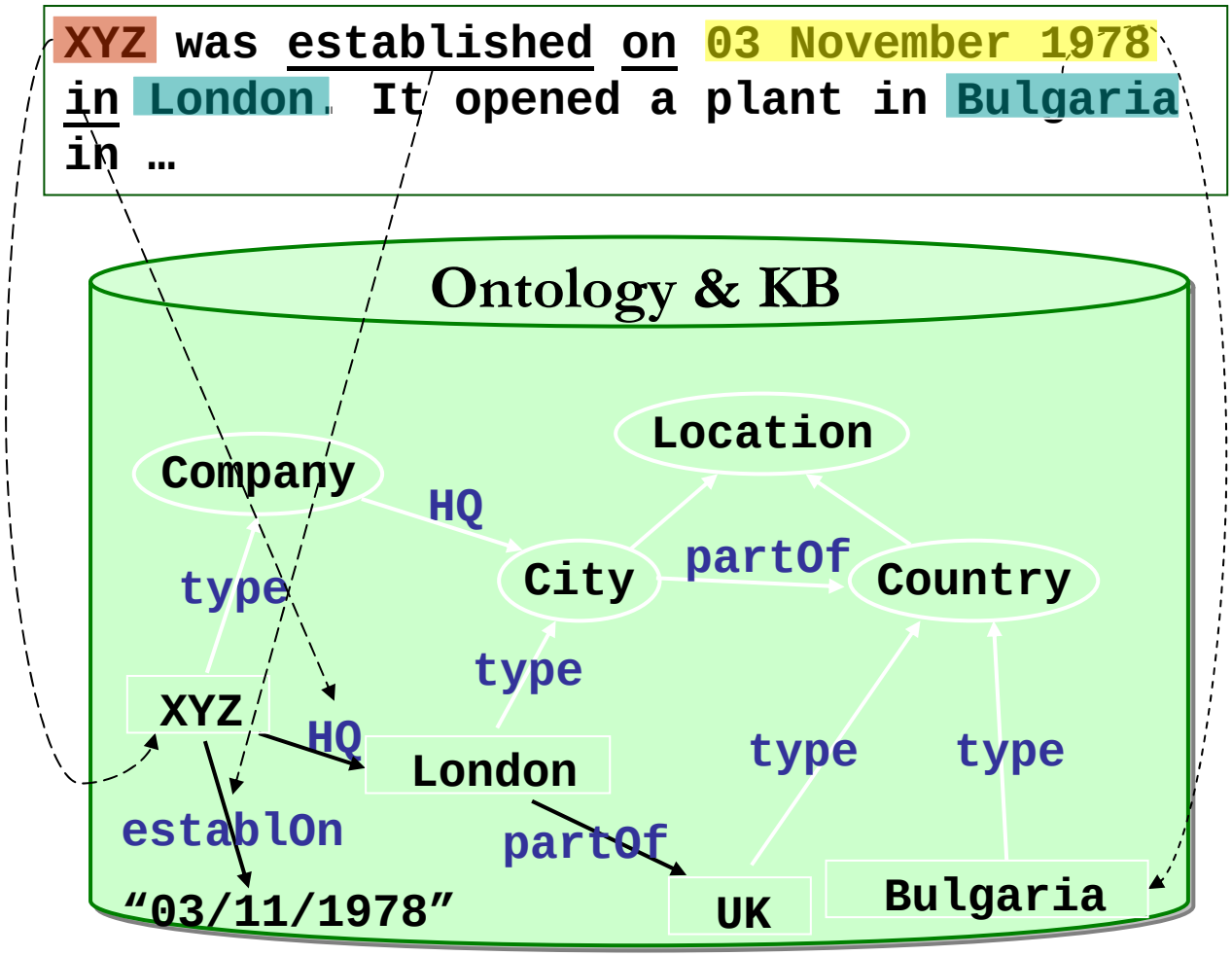
Viewing data: identifying patterns in corpora

- ANNIC – ANNotations In Context
- Provides a keyword-in-context-like interface for identifying annotation patterns in corpora
- Concordancing possible not only over text elements, but over any type of annotation that has been added to the text
- {Token.string="Baroness"}{Token.string="Thatcher"}: find all occurrences of the string "Baroness Thatcher"
- {Token.category="NN"} {Token.category="NN"}:
  Find all noun-noun combinations
- {Person}: find all Person annotations

ANNIC example



Semantic Annotation: Example



Evaluation metrics and tools

- Evaluation metrics mathematically define how to measure the system's performance against human-annotated gold standard
- Scoring program implements the metric and provides performance measures
  - for each document and over the entire corpus
  - for each type of semantic class
  - may also evaluate changes over time
- A gold standard reference set needs to be provided – this may be time-consuming to produce
- Visualisation tools show the results graphically and enable easy comparison

Conclusion

- Gate is a text mining tool that contains synchronic language analysis modules that are of direct interest to diachronic linguistic research;
- As a platform architecture, it can incorporate any text processing algorithm;
- Therefore it allows extension and adaptation to any type of research specific historical text mining
- It enables integration of research effort and data sharing.
- Manuals, tutorials and support available.