The University of Sheffield
Computer Science Department
http://www.dcs.shef.ac.uk

The Sheffield NLP group
http://nlp.shef.ac.uk

GATE — General Architecture for Text Engineering

http://gate.ac.uk/

GATE is an architecture, framework and development environment for R&D in language and knowledge technologies. The system is free software under the GNU Library licence, and includes:

- stable, robust and scaleable language processing tools;
- integration with Semantic Web infrastructure such as Protégé, Sesame and KAON and support for XML, RDF(S), DAML+OIL, etc.;
- components for Information Retrieval, annotation, evaluation, visualisation, persistence, treebanking, enhanced Unicode support, machine learning; Digital Libraries, etc. etc.

GATE and its NLE components have been successfully used in a large number of research projects and commercial applications, including Information Extraction in multiple languages, from multimedia sources, and for multiple tasks and clients. The system has been in development at the University of Sheffield since 1995.

A partial snapshot of current GATE users includes:

- Pearson Educational, UK;
- Merck KgAa, Germany;
- Canon Europe, UK;
- Knight Ridder, US;
- BBN Technologies, US;
- Reuters Inc., US;
- OntoText / Sirma AI Ltd., Bulgaria;
- Resco AB, Sweden/Finland/Germany;
- Glaxo Smith Kline PLC, UK;
- the American National Corpus project, US;
- the Perseus Digital Library project, Tufts University, US;
- Imperial College, London, the University of Manchester, Queen Mary College, UMIST, the University of Karlsruhe, Vassar College, ISI (the University of Southern California) and a large number of other Universities around the world.

Education

Http://gate.ac.uk/teaching.html

Postgraduates in locations as diverse as Sofia, Copenhagen and London are using the system in order to avoid having to write simple things like sentence splitters from scratch, and to enable visualisation and management of data. For example, Partha Lal at Imperial College is developing a summarisation system based on GATE and ANNIE. (His site includes the URL of his components; give GATE the URL and it will load his software over the network.) Marin Dimitrov of the University of Sofia has produced an anaphora resolution system for GATE.

Commercial Applications

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

- **Engineered** to a high standard for deployment in commercial applications software.
- Based on components, mobile code and internet-based distribution.
- 100% **Java** with support for **XML**, **HTML** and **relational databases** (**Oracle, PostgreSQL**).
- Large cross-platform **regression test suite**.
- IE software quality-controlled by the rigorous application of **quantitative evaluation** metrics that ensure that the behaviour of our systems is **predictable** and **robust.**

Scientific Research

Http://gate.ac.uk/science.html

- **Repeatability**: making it easier to repeat comparable experiments across different sites.
- **Quantitative evaluation:** built-in quantitative metrics, e.g. precision and recall.
- **Collaboration:** easy software integration and porting.
- **Reuse not reinvention**: reusing results does not require learning fresh installation and usage conventions for every tool.

GATE has been funded by the EPSRC
http://www.epsrc.ac.uk/

Contact:

Prof. **Yorick Wilks,** or
Dr. **Hamish Cunningham**
Senior Research Scientist
Department of Computer Science
University of Sheffield, UK.
Email: hamish@dcs.shef.ac.uk
Web: http://gate.ac.uk/hamish
Phone: +44 114 222 1891

Portable Information Extraction

http://gate.ac.uk/ie/

GATE is distributed with a **Portable Information Extraction** component set called ANNIE, useable in many different applications:

- Copes seamlessly with documents in many different formats, from badly-spelled lower case email messages to structured XML or HTML pages to newswires to18th century court reports.
- Able to process large data volumes without crashing and at high speed, scaling from desktops to very mainframes and clusters.
- System developers can adapt the system to new circumstances with a minimum of effort.
- Users can adapt the system as far as is possible (some IE tasks cannot be attempted by unskilled users, but where the data is simple end-users can and should be allowed to update the system).
- ANNIE is in use in Romanian, Bulgarian, Greek, Bengali, Spanish, Swedish, German, Italian, French (Arabic, Chinese and Russian next year).
- Integrated with the WEKA machine learning toolkit and the OntoText HMM suite.

MUltiMedia Indexing and Searching environment

Digital Libraries

Http://gate.ac.uk/digilibs.html/

As digital libraries grow in size and coverage, so does the need for automatic content annotation and indexing. GATE's robust and customisable Named Entity recognition and Information Extraction technology has already been used successfully for metadata creation, automatic name and event annotation, indexing, and access. Applications:

- OldBaileyIE required adaption to the non-standard written conventions of Old English in Old Bailey court reports from the 18th Century;
- in MUMIS (Multimedia Indexing and Search) we annotate material in multiple modalities to build a conceptual index of football videos;
- EMILLE focuses on collection and annotation of large corpora in non-indigenous minority languages in the UK (Urdu, Bengali, Sylheti and others).

We are currently looking at GATE for the creation of computational tools for the study of digital collections in cultural heritage languages, e.g. Ancient Greek, Latin.

Multilingual Language Resources

Http://gate.ac.uk/tao/

- Facilities for developing annotated corpora and other Language Resources (LRs).
- Annotation model compatible with the XCES and ATLAS systems; Xschema type model.
- Visualisation and editing tools support trees, chains and flat annotations structures.
- Fully supports multilingual LRs with Unicode facilities beyond default Java:
  - a Unicode editor with input methods for many languages;
  - use of the input methods in all places where text is edited in the GUI;
  - a development kit for implementing input methods;
  - ability to read diverse character encodings.

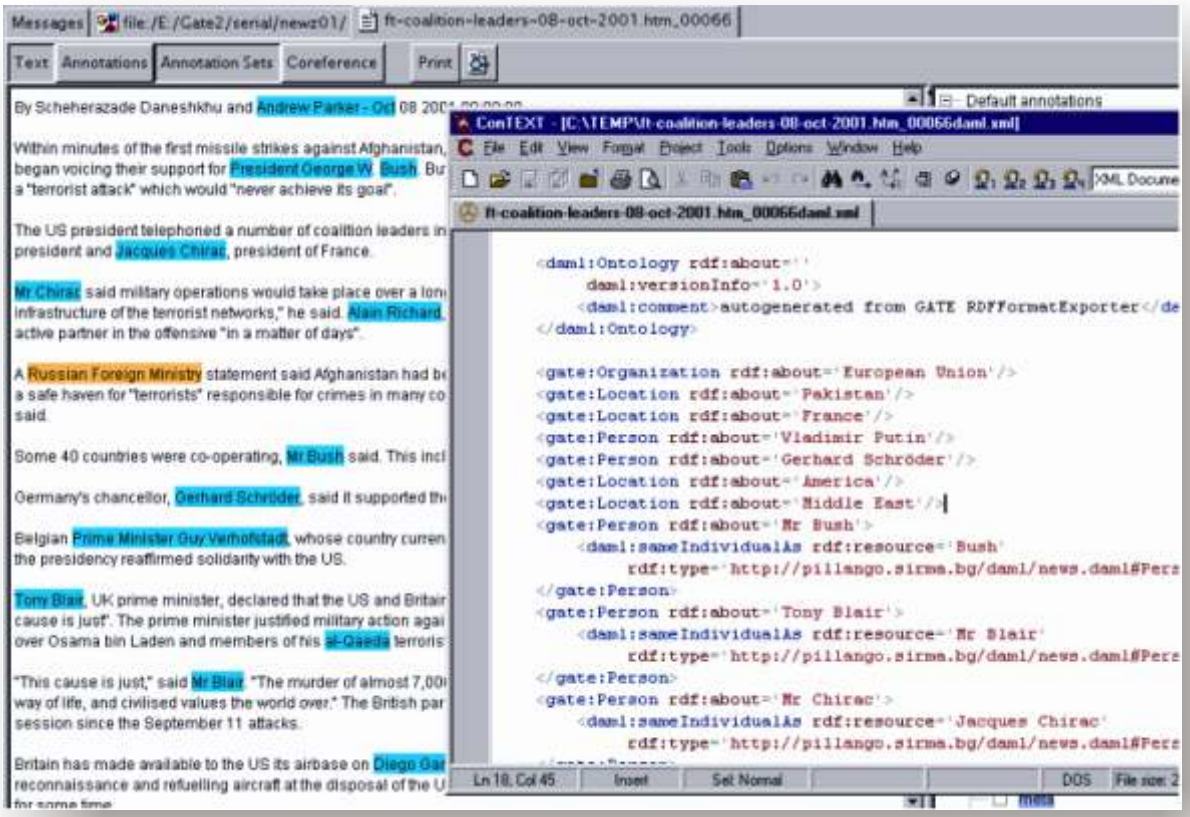
The Semantic Web

Http://gate.ac.uk/semweb.html

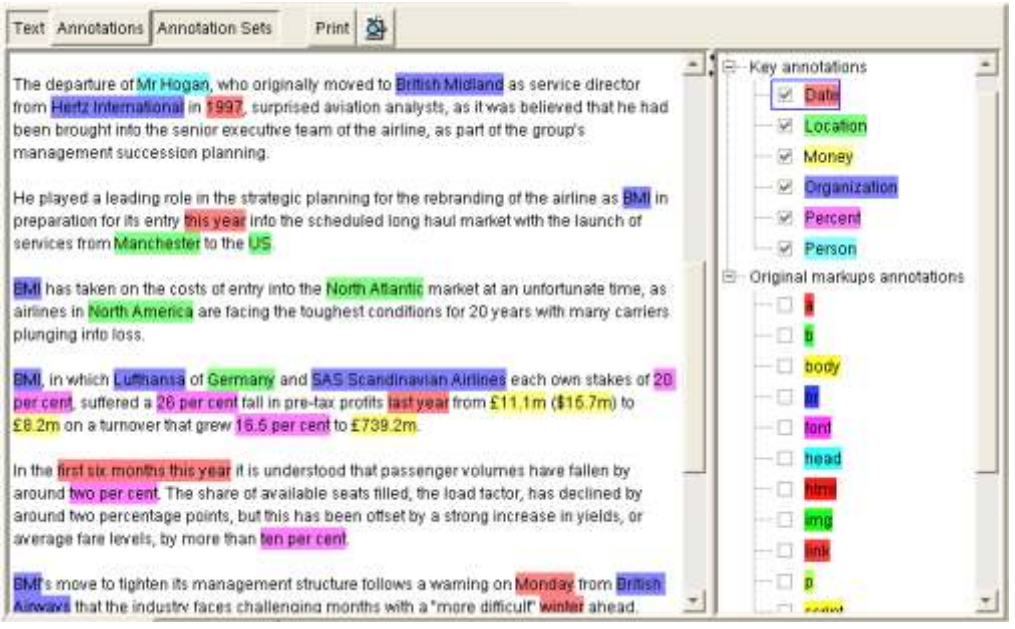
The Semantic Web is adding a machine-tractable layer to the natural language web of HTML. The benefits of success will be many, but the project is currently lacking the critical mass necessary to demonstrate these benefits beyond a few small-scale trial applications. GATE is being used for experiments in automatic and semi-automatic methods for:

- linking web pages to Ontologies using Information Extraction;
- learning and evolving Ontologies via IE and lexical semantic network traversal.

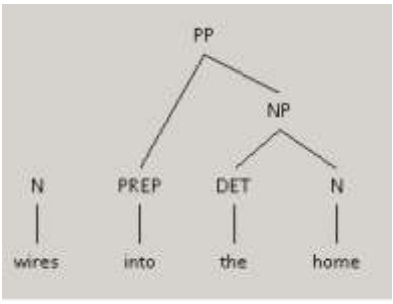
We have also integrated the Protégé Ontology editor with the system. GATE forms the basis of the language technology under development in the UK's Advanced Knowledge Technologies six-year multi-site programme.



Annotations editor/viewer



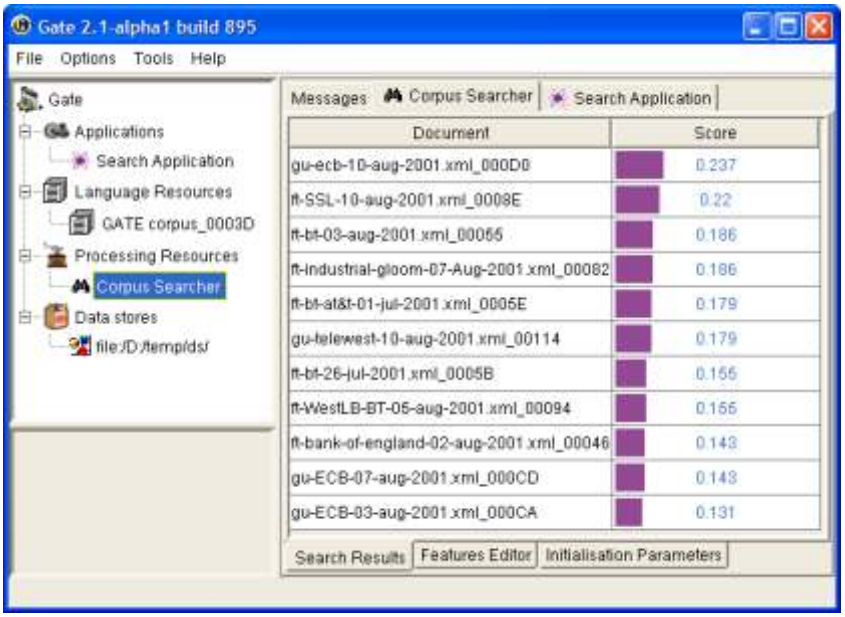
Syntax Tree Editor



Information Retrieval

Http://gate.ac.uk/tao/

- Full-featured Information Retrieval for queries on corpora.
- Documents can be retrieved textual content plus annotations.
- Current implementation based on Lucene; others may be added in future.



Dialogue

Http://gate.ac.uk/dialogue.html/

GATE is being used in the Amities project to produce dialogue processing server components to run in the Galaxy Communicator architecture, which concentrates on distributed processing, hooking together sets of servers and clients that collaborate to hold dialogues with human interlocutors. GATE ignores the issue of distributed processing, and concentrates on facilities for bootstrapping, developing, testing and deploying language processing components. We used GATE to produce a Galaxy Communicator server component, and there seems a natural synergy between the two systems. In future work we would like to more closely integrate GATE with Galaxy Communicator.

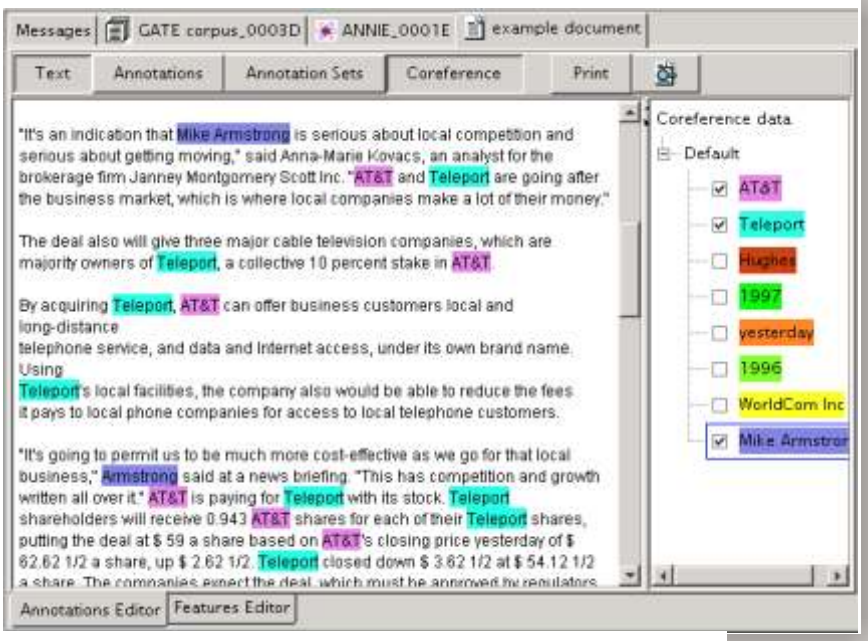
Evaluation

Http://gate.ac.uk/sale/tao/

Two mechanisms for automated performance measurement and visualisation of the results.

- Annotation Diff compares and shows differences on a single document, and produces Precision, Recall, F-measure and Error Rate statistics.
- Regression Test tool tracks system over time over a whole corpus.

Coreference Viewer



GATE has been developed by:

Hamish Cunningham, Valentin Tablan, Kalina Bontcheva, Diana Maynard, Marin Dimitrov,

Cristian Ursu, Oana Hamza, Bobby Popov, Angel Kirilov,Atanas Kiryakov,

Damyan Ognyanov, Horacio Saggion, Mark Leisher, Wim Peters