GATE 2 - IGR Review Report.

Hamish Cunningham,
Yorick Wilks

June 2002

Institute for Language, Speech and Hearing (ILASH), and
Department of Computer Science
University of Sheffield, UK

hamish,yorick@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/~hamish,yorick
http://gate.ac.uk/

1 Introduction

This report reviews the last three years of work on GATE, a General Architecture for Text Engineering, under EPSRC grant GM/M31699. GATE is an infrastructure for support of human language computation work, based on an extendable component model and delivered with a range of free components for common language processing tasks. This project has extended and redeveloped the successful version 1 of GATE, which was used in many R&D projects, PhDs and MScs, for teaching and for a number of commercial applications [Maynard *et al.* 00].

- Section 2 briefly describes the context of the work.
- Section 3 covers the key advances that we have made in this project.
- Section 4 evaluates our work relative to the objectives and plan described in the proposal.
- Section 5 discusses the impact the project has made, and who has used our work.
- Section 6 looks to the future of GATE.

There are two appendices:

- Appendix A lists projects using GATE.
- Appendix B presents an annotated list of publications relating to GATE. The GATE 2 project has led to 4 journal papers, 18 refereed conference papers, 4 refereed workshop papers, 4 technical reports, 1 PhD thesis and 1 MSc thesis.

(**Note:** this report may be read on-line[1] at the GATE web site.)

[1] `http://gate.ac.uk/sale/igr/`

2 Background / Context

The context of this project, computation with human language, encompasses a range of areas spanning science, engineering and commerce:

Computational Linguistics: part of the science of language that uses computation as an investigative tool.

Natural Language Processing: part of the science of computation whose subject matter is data structures and algorithms for human language processing.

Language Engineering: building systems whose cost and outputs are measurable and predictable and therefore appropriate for commercial software development.

(Other valid interpretations of these terms exist – see [Cunningham 99a] for a review.) In this context we may define our work as infrastructure for these related fields as follows:

Software Architecture: macro-level organisational principles for families of systems. In this context the term is also used as infrastructure.

Software Architecture for Language Engineering (SALE): software infrastructure, architecture and development tools for NLP, CL and LE.

GATE is a Software Architecture for Language Engineering.

We began work in this area in the early 1990s [Cunningham 94, Cunningham *et al.* 94]. In 1996 we released version 1 of GATE which was successful mainly as an Information Extraction workbench [Cunningham *et al.* 98b, Stevenson *et al.* 98, Cunningham *et al.* 97, Cunningham *et al.* 96a, Cunningham *et al.* 96b, Gaizauskas *et al.* 96, Cunningham *et al.* 95]. In 1999 we began work on version 2 of GATE, which is the project reported here.

3 Key Advances

We believe that our work over the past three years has make a strong contribution in a number of respects:

Supporting scientific research. GATE contributes to science in four ways:

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

Supporting education. Postgraduates in locations as diverse as Bulgaria, Copenhagen and Surrey 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. GATE is an ideal starting point for student projects on language analysis, as it comes with a set of Information Extraction modules that can be used as a base, and a significant number of PhD students have used GATE in their research. See also [Bontcheva *et al.* 02a] from the ACL 2002 workshop on teaching NLP.

Promoting technology transfer. The following corporates (and a number of SMEs) have used systems based on GATE: GlaxoSmithKline PLC, Reuters PLC, Master Foods NV, British Gas PLC, Merck KGaA. IBM TJ Watson Labs have just announced their intention to use the system.

Defining SALE. [Cunningham 00] and [Cunningham *et al.* 00b] reviewed some hundred papers in the area of infrastructure for language processing work, and derived a set of requirements for such systems that were used to inform the design of GATE 2. We also described means for evaluating this type of system. This element of our work has set SALE on a firm theoretical foundation.

Building GATE version 2. The new version of GATE developed in this project is open source Java software under the GNU library licence, and is a stable, robust, and scalable infrastructure which allows users to focus on LE tasks, while mundane tasks like data storage, format analysis and data visualisation are handled by GATE. The system is bundled with components for language analysis, and is in use for Information Extraction (IE), Information Retrieval (IR), summarisation, dialogue, Semantic Web, Knowledge Technologies and Digital Libraries applications. GATE-based systems have taken part in the TREC (IR), ACE (IE, successor to MUC) and DUC (summarisation) evaluation programmes in recent years.

4 Project Plan Review

This section reviews the outcome of the project relative to the original proposal plan.

The plan (see the 'objectives and research summary' section of the IGR form) mentions these specific items:

Research productivity. Appendix A lists some 20 projects using GATE at the time of writing. Irrespective of the productivity benefits of using the development environment, in each case some basic software that would otherwise need to be written from scratch is reused from GATE. [Cunningham 00] conducted a detailed evaluation of experience with GATE version 1 and found significant productivity gains. We believe that this benefit also exists for the new version.

Information Extraction. GATE is at the core of Sheffield's research on IE, which has performed well in the MUC and ACE evaluations, and encompasses a wide variety of approaches including knowledge-based [Gaizauskas & Wilks 98], GATE has been used to do IE in English, Bulgarian, Romanian, Bengali, Greek, Spanish, Swedish, German, Italian, and French.

Machine Translation. Version 1 of GATE was biased towards language analysis software, and so was inappropriate for MT and Natural Language Generation (NLG) work. We have corrected this bias, and integrated NLG tools with the system. (These components are in use for the AKT project – see below.)

Information Retrieval. We have developed a model for IR tools in with GATE, and integrated the Lucene open source full text retrieval system. This means that GATE corpora can be queried on

their text contents, and also, at an API level, arbitrary features of documents can be indexed (e.g. to allow searches on syntax, or dialogue act).

Dialogue processing. GATE is being used in the US/EU Amities project to produce dialogue processing server components to run in the Galaxy Communicator architecture.

Distributed data model. In GATE language resources such as documents, corpora, lexicons, etc., are stored in a variety of databases (including Oracle and PostgreSQL[2]). These databases are accessed using standard internet protocols such as JDBC, thus allowing the data to reside anywhere on the net.

Multilinguality. GATE 2 is based on Unicode, which solves many but (to our surprise) not all issues about supporting multilinguality. We have added an extendable editing support kit for a number of languages, and a facility for creating font mapping tables to allow conversion to Unicode from diverse encodings. (These facilities are in use in the EMILLE Indic languages project – see below.)

Multi-media documents. We ran a pilot project with the Max Planck Institute on integrating GATE with their EUDICO multimedia annotation tool [Brugman *et al.* 99], and have developed IE for multimedia data as part of the EU MUMIS project [Saggion *et al.* 02b, Saggion *et al.* 02a] on conceptual indexing of football games.

Distributed processing. Here we have deviated from our plan: GATE does not directly support distribution of processing. We decided not to pursue this goal based on the negative experiences of the Calypso project [Zajac 98]. We have, however, implemented distribution of the implementations of processes: GATE processing resources can be loaded over the net given a URL and some XML metadata.

In addition we have worked in a number of areas that were not in the original plan:

The Semantic Web. We have integrated the Protégé ontology editor, added ontologies to GATE's Language Resources model, produced a gazetteer system that rapidly assigns text instances to concepts, and developed a DAML+OIL export for annotating web pages (also in use in the AKT project).

Digital Libraries. We have applied the system to the development of corpora for historical research with the Humanities Research Institute, we have worked on annotation of multimedia data and we are collaborating with Tufts University (Boston) on future projects in annotation of Latin and Greek for humanities research.

Summarisation. GATE has been used for several projects on summarisation [Maynard *et al.* 02c].

In sum, we feel that we have exceeded the contribution that we promised in the project proposal.

5 Research Impact and Benefits to Society

GATE has had a significant impact, and is among the mostly widely used and cited systems in its field. It was specifically cited for its wide usage by a team of international reviewers that produced a document for the EPSRC, IEE and BCS (International Review of UK Research in Computer Science, ed. F.B. Schneider and M. Rodd, June 2001, p. 22). Many R&D workers in the language processing field have used GATE, and since the release of version 2 in early 2002 interest in the system has increased rapidly. Several recent developments are especially significant:

[2]Simple file-based storage is also available.

- The American National Corpus team are adopting GATE for annotation in that project. This will ensure that GATE's support for XCES and other XML standards remains current with the community.
- The ATLAS team at NIST (the US National Institute of Standards and Technology) are making their resources available in GATE. This will make resources using the LDC[3]'s annotation graph available in the system.
- We have integrated Stanford's Protégé ontology editor with GATE, and made ontologies available as resources within the system. Protégé is very widely used in the knowledge technologies field.

These developments indicate that GATE is becoming an accepted part of the infrastructural landscape in language processing. We believe that the benefits of the system to the community that spawned it will continue to grow.

GATE is currently in use in many research projects – see appendix A.

6 Future Work

Since releasing version 2 we have been commited to backwards compatibility, and we no longer make changes to the core of the system that may disrupt our existing user base. Because the system is component-based we are still able to extend its capabilities. In future we plan to develop the system in a number of directions:

- Adoption of the RAGS (Reference Architecture for Generation Systems) model for NLG components integrated with GATE. RAGS is a model of the data structures involved in NLG, and will benefit from GATE's advanced grapical tools, GATE's model of NLP software as dynamic components sets loaded over a network, and GATE's facilities for storing and indexing arbitrary data.
- Development of a shared server accessible over Janet supporting experiments on resources in the TEI, (X)CES and ATLAS formats. The facility to upload mobile code to a central server and receive the results of a parallelised application process.
- Galaxy Communicator (from Mitre Corporation, supported by DARPA), like the Verbmobil ICE infrastructure (from DFKI), is a software architecture oriented to the distribution and asynchronous execution aspects of dialog processing. GATE is oriented on data visualisation and management, and component-based computing. We will implement a bridge between the two models, with benefits to both infrastructures.
- Protégé has become the de facto standard for Ontological data creation and visualisation, and key component of the story of the Semantic Web. Our facilities for Protégé and others (Ontological Gazetteer, Wordnet, DAML+OIL export) are designed to ensure GATE's contribution to the development of the web.
- Continuation of our leading work on IE, with open source results (our ANNIE system is one of few open source, well-documented IE systems available).
- Maintenance and support of GATE for its user community in the UK and elsewhere.

[3]Linguistic Data Consortium.

Appendices

A Projects Using GATE

This appendix lists most of the projects (that we know of[4]) that are currently using GATE.

Enactable Models[5] *Middlesex University.* Building a summarisation system based on discourse structure.

DUC *Imperial College, London.* Building a summarisation system to be entered in the Document Understanding Conference (DUC) evaluation.

Parallel IE *Merck kGaA, Darmstadt.* Information Extraction on a Linux cluster for bio-medical text mining and indexing.

GROK/OpenNLP *University of Edinburgh.* Integration of GATE with the GROK/OpenNLP project – a library of NLP components including support for parsing and various preprocessing tasks.

Medline Analysis *Institute for Medical Informatics and Biometry, University of Rostock, Germany.* Analysing MEDLINE abstracts to extract causal functional relations, which are essential for the construction of genetic networks, as a step towards characterisation of diseases.

Database technology[6] *Birkbeck College, London.* Using IE to enhance the support for text in database technology.

OntoWeb SIG5, Language Technology in Ontology Development and Use[7] *CNTS Antwerp, DFKI Saarbrucken, FZI Karlsruhe, Language and Computing NV, University of Sheffield, Ontoprise, Sirma AI Ltd.* The OntoWeb network of Excellence in semantic web technologies, language technology Special Interest Group.

EMILLE[8] *University of Lancaster, University of Sheffield.* EMILLE [McEnery *et al.* 00] is building a 63 million word electronic corpus of South Asian languages, especially those spoken in the UK.

Question Answering *University of Sheffield.* The QA project for building a question answering system for entry into TREC.

ArtEquAkt *University of Southampton, University of Sheffield.* This e-science project, producing composite descriptions of cultural artefacts and figures (e.g. Rembrandt) from diverse web pages, uses GATE-based Natural Language Generation system. ArtEquAkt is a collaboration between the Equator wearable computing project and the AKT Knowledge Technologies project.

htlinkhttp://www.cs.man.ac.uk/ai/MultiFlora/Multiflora II *University of Manchester, University of Sheffield.* An e-science bioinformatics project for biodiversity support.

MiAKT *University of Southampton, University of Sheffield, Open University, Oxford University, Guy's Hospital, King's College London.* The MiAKT e-science project, which involves collaborative problem solving environments in Medical Informatics, using knowledge services provided by the e-Science grid infrastructure.

MUSE[9] *University of Sheffield.* Named entity recognition from diverse text types and genres.

[4] Our experience with version 1 was that we frequently did not find out about projects using the system until years later. 'GATE' being a common word, it is relatively difficult to search the web for projects.

MUMIS[10] *CTIT (Netherlands), University of Sheffield, University of Nijmegen (Netherlands), DFKI (Germany), Max Planck Institut fr Psycholinguistik (Germany), ESTEAM (Sweden), VDA (Netherlands).* Automatic creation of indexes into multimedia programme material, using data from several sources and several languages, in the domain of football.

SOCIS[11] *University of Sheffield, University of Surrey.* Integration of knowledge acquisition, information extraction, image processing and speech recognition technologies in the domain of police crime reports.

AKT[12] *University of Anerdeen, University of Edinburgh, Open University, University of Sheffield, University of Southampton.* Advanced Knowledge Technologies.

OldBaileyIE *University of Sheffield.* Named entity recognition on 17th century Old Bailey Court reports.

HealthIE *University of Sheffield.* Summarisation of information from company reports to generate statistics about the level of compliance with Health and Safety recommendations and legislation.

AMITIES[13] *University of Sheffield, CNRS-LIMSI, GE Service Centre GMBH, VECSYS, VIEL and CIE, State University of New York DUke UNiversity, GE Research and Development.* Building empirically induced dialogue processors to support multilingual human-computer interaction.

CLEF *University of Manchester, University of Cambridge, University of Sheffield, University College London, Royal Marsden NHS Trust, University of Brighton.* Clinical E-science Framework. Building on E-Science technology to embed a full information cycle within practical clinical systems, building tools to integrate patient information from text and images, and linking clinical and genomic research.

MyGRID[14] *University of Manchester, University of Newcastle, University of Nottingham, University of Sheffield, University of Southampton, IT Innovation Centre, European Bioinformatics Institute.* Extending the GRID framework of distributed conputing by producing a virtual laboratory bench that will support the life sciences community and make use of complex distributed resources.

B Annotated List of Publications

This appendix gives an annotated list of publications on GATE and projects that use it (some of which are available on-line[15]). Since 1999 the GATE 2 project has led to 4 journal papers, 18 refereed conference papers, 4 workshop papers, 4 technical reports, 1 PhD thesis and 1 MSc thesis.

[**Cunningham** *et al.* **02a**] (ACL 2002) describes the GATE framework and graphical development environment as a tool for robust NLP applications.

[**Cunningham** *et al.* **02b**] (web site) is a 170-page user guide to GATE.

[**Bontcheva** *et al.* **02b**] (NLIS 2002) discusses how GATE can be used to create HLT modules for use in information systems.

[**Tablan** *et al.* **02**] (LREC 2002) describes GATE's enhanced Unicode support.

[**Maynard** *et al.* **02c**] (ACL 2002 Summarisation Workshop) describes using GATE to build a portable IE-based summarisation system in the domain of health and safety.

[15] `http://gate.ac.uk/gate/doc/papers.html`

[**Bontcheva** *et al.* **02a**] (ACl 2002 Workshop) describes how GATE can be used as an environment for teaching NLP, with examples of and ideas for future student projects developed within GATE.

[**Maynard** *et al.* **02d**] (Nordic Language Technology) describes various Named Entity recognition projects developed at Sheffield using GATE.

[**Maynard** *et al.* **02a**] (AIMSA 2002) describes the adaptation of the core ANNIE modules within GATE to the ACE (Automatic Content Extraction) tasks.

[**Maynard** *et al.* **02b**] (JNLE) describes robustness and predictability in LE systems, and presents GATE as an example of a system which contributes to robustness and to low overhead systems development.

[**Bontcheva** *et al.* **02c**], [**Dimitrov** *et al.* **02**] and [**Dimitrov 02**] (TALN 2002, DAARC 2002, MSc thesis) describe the shallow named entity coreference modules in GATE: the orthomatcher which resolves pronominal coreference, and the pronoun resolution module.

[**Pastra** *et al.* **02**] (LREC 2002) discusses the feasibility of grammar reuse in applications using ANNIE modules.

[**Baker** *et al.* **02**] (LREC 2002) report results from the EMILLE Indic languages corpus collection and processing project.

[**Saggion** *et al.* **02c**] and [**Saggion** *et al.* **02a**] (LREC 2002, SPLPT 2002) describes how ANNIE modules have been adapted to extract information for indexing multimedia material.

[**Maynard** *et al.* **01**] (RANLP 2001) discusses a project using ANNIE for named-entity recognition across wide varieties of text type and genre.

[**Cunningham 00**] (PhD thesis) defines the field of Software Architecture for Language Engineering, reviews previous work in the area, presents a requirements analysis for such systems (which was used as the basis for designing GATE version 2), and evaluates the strengths and weaknesses of GATE version 1.

[**Cunningham 02**] (Computers and the Humanities) describes the philosophy and motivation behind the system, describes GATE version 1 and how well it lived up to its design brief.

[**McEnery** *et al.* **00**] (Vivek) presents the EMILLE project[16] in the context of which GATE's Unicode support for Indic languages has been developed.

[**Cunningham** *et al.* **00d**] and [**Cunningham 99b**] (technical reports) document early versions of JAPE (superceded by the present document).

[**Cunningham** *et al.* **00a**], [**Cunningham** *et al.* **98a**] and [**Peters** *et al.* **98**] (OntoLex 2000, LREC 1998) presents GATE's model of Language Resources, their access and distribution.

[**Maynard** *et al.* **00**] (technical report) surveys users of GATE up to mid-2000.

[**Cunningham** *et al.* **00c**] and [**Cunningham** *et al.* **99**] (COLING 2000, AISB 1999) summarise experiences with GATE version 1.

[**Cunningham** *et al.* **00b**] (LREC 2000) taxonomises Language Engineering components and discusses the requirements analysis for GATE version 2.

[**Bontcheva** *et al.* **00**] and [**Brugman** *et al.* **99**] (COLING 2000, technical report) describe a prototype of GATE version 2 that integrated with the EUDICO multimedia markup tool[17] from the Max Planck Institute.

[16] `http://www.emille.lancs.ac.uk/`

[17] `http://www.mpi.nl/world/tg/lapp/eudico/eudico.html`

[Gambäck & Olsson 00] (LREC 2000) discusses experiences in the Svensk project, which used GATE version 1 to develop a reusable toolbox of Swedish language processing components.

[Cunningham 99a] (JNLE) reviewed and synthesised definitions of Language Engineering.

References

[Baker *et al.* 02]
P. Baker, A. Hardie, T. McEnery, H. Cunningham, and R. Gaizauskas. EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In *Proceedings of 3rd Language Resources and Evaluation Conference (LREC'2002)*, pages 819–825, 2002.

[Bontcheva *et al.* 00]
K. Bontcheva, H. Brugman, A. Russel, P. Wittenburg, and H. Cunningham. An Experiment in Unifying Audio-Visual and Textual Infrastructures for Language Processing R&D. In *Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*, Luxembourg, 2000. `http://gate.ac.uk/`.

[Bontcheva *et al.* 02a]
K. Bontcheva, H. Cunningham, V. Tablan, D. Maynard, and O. Hamza. Using GATE as an Environment for Teaching NLP. In *ACL Workshop on Effective Tools and Methodologies in Teaching NLP*, 2002.

[Bontcheva *et al.* 02b]
K. Bontcheva, H. Cunningham, V. Tablan, D. Maynard, and H. Saggion. Developing Reusable and Robust Language Processing Components for Information Systems using GATE. In *3rd International Workshop on Natural Language and Information Systems (NLIS'2002)*, Aix-en-Provence, France, 2002. IEEE Computer Society Press. To appear.

[Bontcheva *et al.* 02c]
K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham. Shallow Methods for Named Entity Coreference Resolution. In *Chaînes de références et résolveurs d'anaphores, workshop TALN 2002*, Nancy, France, 2002. To appear.

[Brugman *et al.* 99]
H. Brugman, K. Bontcheva, P. Wittenburg, and H. Cunningham. Integrating Multimedia and Textual Software Architectures for Language Technology. Technical report MPI-TG-99-1, Max-Planck Institute for Psycholinguistics, Nijmegen, Netherlands, 1999.

[Cunningham 94]
H. Cunningham. Support Software for Language Engineering Research. Technical Report 94/05, Centre for Computational Linguistics, UMIST, Manchester, 1994.

[Cunningham 99a]
H. Cunningham. A Definition and Short History of Language Engineering. *Journal of Natural Language Engineering*, 5(1):1–16, 1999.

[Cunningham 99b]
H. Cunningham. JAPE: a Java Annotation Patterns Engine. Research Memorandum CS–99–06, Department of Computer Science, University of Sheffield, May 1999.

[Cunningham 00]
H. Cunningham. *Software Architecture for Language Engineering.* Unpublished PhD thesis, University of Sheffield, 2000. `http://gate.ac.uk/sale/thesis/`.

[Cunningham 02]
H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002.

[Cunningham *et al.* 94]
H. Cunningham, M. Freeman, and W. Black. Software Reuse, Object-Oriented Frameworks and Natural Language Processing. In *New Methods in Language Processing (NeMLaP-1), September 1994*, Manchester, 1994. (Re-published in book form 1997 by UCL Press).

[Cunningham *et al.* 95]
H. Cunningham, R. Gaizauskas, and Y. Wilks. A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D. Technical Report CS–95–21, Department of Computer Science, University of Sheffield, 1995. `http://xxx.lanl.gov/-abs/cs.CL/9601009`.

[Cunningham *et al.* 96a]
H. Cunningham, Y. Wilks, and R. Gaizauskas. GATE – a General Architecture for Text Engineering. In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*, Copenhagen, August 1996.

[Cunningham *et al.* 96b]
H. Cunningham, Y. Wilks, and R. Gaizauskas. New Methods, Current Trends and Software Infrastructure for NLP. In *Proceedings of the Conference on New Methods in Natural Language Processing (NeMLaP-2)*, Bilkent University, Turkey, September 1996. `http://xxx.lanl.gov/-abs/cs.CL/9607025`.

[Cunningham *et al.* 97]
H. Cunningham, K. Humphreys, R. Gaizauskas, and Y. Wilks. Software Infrastructure for Natural Language Processing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, March 1997. `http://xxx.lanl.gov/abs/cs.CL/9702005`.

[Cunningham *et al.* 98a]
H. Cunningham, W. Peters, C. McCauley, K. Bontcheva, and Y. Wilks. A Level Playing Field for Language Resource Evaluation. In *Workshop on Distributing and Accessing Lexical Resources at Conference on Language Resources Evaluation, Granada, Spain*, 1998.

[Cunningham *et al.* 98b]
H. Cunningham, M. Stevenson, and Y. Wilks. Implementing a Sense Tagger within a General Architecture for Language Engineering. In *Proceedings of the Third Conference on New Methods in Language Engineering (NeMLaP-3)*, pages 59–72, Sydney, Australia, 1998.

[Cunningham *et al.* 99]
H. Cunningham, R. Gaizauskas, K. Humphreys, and Y. Wilks. Experience with a Language Engineering Architecture: Three Years of GATE. In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP*, Edinburgh, April 1999. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

[Cunningham *et al.* 00a]
H. Cunningham, K. Bontcheva, W. Peters, and Y. Wilks. Uniform language resource access and distribution in the context of a General Architecture for Text Engineering (GATE). In *Proceedings of the Workshop on Ontologies and Language Resources (OntoLex'2000)*, Sozopol, Bulgaria, September 2000. `http://gate.ac.uk/sale/ontolex/ontolex.ps`.

[Cunningham *et al.* 00b]
H. Cunningham, K. Bontcheva, V. Tablan, and Y. Wilks. Software Infrastructure for Language Resources: a Taxonomy of Previous Work and a Requirements Analysis. In *Proceedings of the*

2nd International Conference on Language Resources and Evaluation (LREC-2), Athens, 2000. http://gate.ac.uk/.

[Cunningham *et al.* 00c]
H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and Y. Wilks. Experience of using GATE for NLP R&D. In *Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*, Luxembourg, 2000. http://gate.ac.uk/.

[Cunningham *et al.* 00d]
H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November 2000.

[Cunningham *et al.* 02a]
H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[Cunningham *et al.* 02b]
H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu. *The GATE User Guide*. http://gate.ac.uk/, 2002.

[Dimitrov 02]
M. Dimitrov. *A Light-weight Approach to Coreference Resolution for Named Entities in Text*. MSc Thesis, University of Sofia, Bulgaria, 2002. http://www.ontotext.com/ie/thesis-m.pdf.

[Dimitrov *et al.* 02]
M. Dimitrov, K. Bontcheva, H. Cunningham, and D. Maynard. A Light-weight Approach to Coreference Resolution for Named Entities in Text. In *Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lisbon, 2002. forthcoming.

[Gaizauskas & Wilks 98]
R. Gaizauskas and Y. Wilks. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105, 1998.

[Gaizauskas *et al.* 96]
R. Gaizauskas, H. Cunningham, Y. Wilks, P. Rodgers, and K. Humphreys. GATE – an Environment to Support Research and Development in Natural Language Engineering. In *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-96)*, Toulouse, France, October 1996.

[Gambäck & Olsson 00]
B. Gambäck and F. Olsson. Experiences of Language Engineering Algorithm Reuse. In *Second International Conference on Language Resources and Evaluation (LREC)*, pages 155–160, Athens, Greece, 2000.

[Maynard *et al.* 00]
D. Maynard, H. Cunningham, K. Bontcheva, R. Catizone, G. Demetriou, R. Gaizauskas, O. Hamza, M. Hepple, P. Herring, B. Mitchell, M. Oakes, W. Peters, A. Setzer, M. Stevenson, V. Tablan, C. Ursu, and Y. Wilks. A Survey of Uses of GATE. Technical Report CS–00–06, Department of Computer Science, University of Sheffield, 2000.

[Maynard *et al.* 01]
D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named Entity Recognition from Diverse Text Types. In *Recent Advances in Natural Language Processing 2001 Conference*, Tzigov Chark, Bulgaria, 2001.

[Maynard *et al.* 02a]
D. Maynard, H. Cunningham, K. Bontcheva, and M. Dimitrov. Adapting A Robust Multi-Genre NE System for Automatic Content Extraction. In *The Tenth International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2002)*, 2002.

[Maynard *et al.* 02b]
D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 2002. forthcoming.

[Maynard *et al.* 02c]
D. Maynard, K. Bontcheva, H. Saggion, H. Cunningham, and O. Hamza. Using a text engineering framework to build an extendable and portable IE-based summarisation system. In *Proceedings of the ACL Workshop on Text Summarisation*, 2002.

[Maynard *et al.* 02d]
D. Maynard, H. Cunningham, and R. Gaizauskas. Named entity recognition at sheffield university. In H. Holmboe, editor, *Nordic Language Technology – Arbog for Nordisk Sprogtechnologisk Forskningsprogram 2002-2004*, pages 141–145. Museum Tusculanums Forlag, 2002.

[McEnery *et al.* 00]
A. McEnery, P. Baker, R. Gaizauskas, and H. Cunningham. EMILLE: Building a Corpus of South Asian Languages. *Vivek, A Quarterly in Artificial Intelligence*, 13(3):23–32, 2000.

[Pastra *et al.* 02]
K. Pastra, D. Maynard, H. Cunningham, O. Hamza, and Y. Wilks. How feasible is the reuse of grammars for named entity recognition? In *Proceedings of 3rd Language Resources and Evaluation Conference*, 2002.

[Peters *et al.* 98]
W. Peters, H. Cunningham, C. McCauley, K. Bontcheva, and Y. Wilks. Uniform Language Resource Access and Distribution. In *Workshop on Distributing and Accessing Lexical Resources at Conference on Language Resources Evaluation*, Granada, Spain, 1998.

[Saggion *et al.* 02a]
H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, C. Ursu, O. Hamza, and Y. Wilks. Access to Multimedia Information through Multisource and Multilanguage Information Extraction. In *7th Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, Stockholm, Sweden, 2002.

[Saggion *et al.* 02b]
H. Saggion, H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks. Extracting Information for Automatic Indexing of Multimedia Material. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 669–676, Las Palmas, Gran Canaria, Spain, 2002.

[Saggion *et al.* 02c]
H. Saggion, H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks. Extracting information for information indexing of multimedia material. In *Proceedings of 3rd Language Resources and Evaluation Conference (LREC'2002)*, 2002.

[Stevenson *et al.* 98]
M. Stevenson, H. Cunningham, and Y. Wilks. Sense tagging and language engineering. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98)*, pages 185–189, Brighton, U.K., 1998.

[Tablan *et al.* 02]
V. Tablan, C. Ursu, K. Bontcheva, H. Cunningham, D. Maynard, O. Hamza, T. McEnery, P. Baker, and M. Leisher. A unicode-based environment for creation and use of language resources. In *Proceedings of 3rd Language Resources and Evaluation Conference*, 2002.

[Zajac 98]
R. Zajac. Reuse and Integration of NLP Components in the Calypso Architecture. In *Workshop on Distributing and Accessing Linguistic Resources*, pages 34–40, Granada, Spain, 1998. `http://www.dcs.shef.ac.uk/~hamish/dalr/`.