Extracting the Science from Scientific Publications

Ann Copestake, Simone Teufel, Peter Murray-Rust and Andy Parker.

This is an edited version of the original project proposal.

This proposal is for a collaboration between three groups in Cambridge. The Computer Laboratory (CL) group provides expertise in Natural Language Processing (NLP), the Chemistry group brings in a user-base as well as experience in application of NLP technology to eScience goals, while the Cambridge eScience Centre (CesC) will ensure scalability and integration of the technology. The project also involves three publishers as partners, who will support the project by supplying large corpora of scientific papers, providing informal feedback and facilitating dissemination activities: the Royal Society of Chemistry, Nature Publishing Group and the International Union of Crystallography.

Project objectives:

1. To develop a natural-language oriented markup language which enables the tight integration of partial information from a wide variety of language processing tools, while being compatible with GRID and Web protocols and having a sound logical basis consistent with Semantic Web standards.

2. To use this language as a basis for robust and extensible extraction of information from scientific texts.

3. To model scientific argumentation and citation purpose in order to support novel modes of information access.

4. To demonstrate the applicability of this infrastructure in a real-world eScience environment by developing technology for Information Extraction and ontology construction applied to Chemistry texts.

1 Background

Our long-term aim is to build dynamic, flexible and expandable natural language processing (NLP) infrastructure which will support applications in eScience. We will show that autonomous, adaptive methods, based on NLP techniques, can be used to mine the primary literature and other text to build an evolving knowledge base for eScience. The fundamental challenge for development of robust, distributed NLP is not interchange protocols at the level of Grid and Web communications but rather the development of a representation which is compatible with these protocols, but which also enables the tight integration of partial information from a wide variety of language processing tools and has a sound logical basis compatible with Semantic Web standards. A central theme of this proposal is the development of a natural-language oriented markup language which meets these criteria. Robust Minimal Recursion Semantics (RMRS) is an application-independent representation which captures the information that comes from the syntax and morphology of natural language.

We will demonstrate the utility of this approach in tasks involving knowledge extraction from text. Most existing Information Extraction (IE) technology is based on relatively shallow processing of texts to directly instantiate domain-specific templates or databases. However, for each new type of information, a hand-crafted system or an extensive manually-created training corpus is required. In contrast, we propose a layered architecture using IE technology that takes the RMRS markup, rather than text, as a starting point. This approach provides a methodology for incrementally incorporating deeper NLP techniques as they evolve in order to enhance IE performance. It also allows us to extract much richer and more varied information from texts than is possible with existing techniques, for instance scientific argumentation structure. Knowledge of the overall discourse structure of the scientific text and interpretation of citation context can enhance human browsing and support more fine-grained searches in the literature.

As discussed in more detail below, preliminary results from experiments with RMRS have demonstrated the feasibility of this approach and shown how several types of existing NLP technology can produce RMRS markup. Here we propose to refine and extend the RMRS approach and to demonstrate its applicability for eScience. Development of this research to its full potential is an ambitious long-term goal, but it has application in the short-term as well.

In this project we will develop a practical tool, the Chemists' Amanuensis, which supports knowledge base acquisition, ontology construction and free-style browsing. This will aid researchers, both in mining the existing scientific record and in supporting authoring of e-publications with appropriate links and annotation. It will also help the larger community, such as publishers and government organisations, to navigate the literature and to annotate existing texts. A key aim is that the initial technology be usable within two years of the start of the project, but be successively augmented as more powerful NLP techniques are deployed.

Our methods are applicable across scientific domains but we here concentrate on Chemistry as a starting point. Peter Murray-Rust's group has already made much progress in utilising Chemistry literature as a knowledge base. They have developed domain-specific processing tools which will integrate well with the existing

NLP technology developed by the CL group. In particular, Chemical Markup Language (CML) and associated technology greatly enhance the feasibility of processing Chemistry texts. Furthermore, there is a huge potential user base, not only in Chemistry itself but in other domains which depend on it: life sciences, health care, materials, nanotechnology, etc. Success in Chemistry will therefore transfer to other disciplines. In contrast to most data-mining projects in biology, which involve relatively specific types of information, Chemistry is much broader and will therefore act as a better test of the new technology. Finally, our partners will provide us with corpora of tens of thousands of papers for the project.

Overall, the goals of this proposal are:

- To develop an NL markup language which will act as a platform for extraction of information.
- To develop IE technology and core ontologies for use by publishers, researchers, readers, vendors, and regulatory organisations.
- To model scientific argumentation and citation purpose in order to support novel modes of information access.
- To demonstrate the applicability of this infrastructure in a real-world eScience environment.

2 Programme and Methodology

To analyse text structure, we will use general purpose NLP techniques combined with specific algorithms for chemistry, combined via RMRS. This allows for different levels of linguistic processing. RMRS is an extension of the Minimal Recursion Semantics (MRS: Copestake et al, 1995 and in press) approach which is well established in 'deep' processing in NLP. By 'deep', we mean systems which use very precise and detailed grammars of natural languages to analyse and generate. MRS is compatible with RMRS but RMRS can also be used with shallow processing techniques, such as part-of-speech tagging, noun phrase chunking and stochastic parsers which operate without detailed lexicons. Shallow processing has the advantage of being more robust and faster, but is less precise: RMRS output from the shallower systems is less fully specified than the output from the deeper systems, but in principle fully compatible.

The advantage of this approach is that application algorithms can be developed which operate on RMRS output regardless of the system which produces it. In circumstances where deep parsing can be successfully applied, a detailed RMRS can be produced, but when resource limitations (in processing capability or lexicon availability, for instance) preclude this, the system can back-off to RMRSs produced by shallower analysers. Different analysers can be flexibly combined: for instance shallow processing can be used as a preprocessor for deep analysis to provide structures for unknown words, to limit the search space or to identify regions of the text which are of particular interest. Conversely, RMRS structures from deep analysis can be further instantiated by anaphora resolution, word sense disambiguation and other techniques. Thus RMRS is used as the common integration language to enable flexible combinations of resources, which has not been previously possible.

Example RMRS output for a POS tagger and a deep parser for the sentence *the mixture was allowed to warm* is shown below:

```
Deep processing        POS tagger
prpstn_m_rel(h1,h5)
  PSV(h1,x3)
  qeq(h5,h11)
_the_q(h6,x3)          _the_q(h1,x2)
  RSTR(h6,h8)
  BODY(h6,h7)
_mixture_n(h9,x3)      _mixture_n(h3,x4)
  ARG1(h9,u10)
_allow_v_1(h11,e2)     _allow_v (h5,e6)
  ARG1(h11,u12)
  ARG2(h11,x3)
  ARG3(h11,h13)
prpstn_m_rel(h13,h14)
  qeq(h14,h17)
_warm_v(h17,e18)       _warm_v(h7,e8)
  ARG1(h17,x3)
```

We cannot describe RMRS in detail here, but note the following points:

- RMRSs consist of 'flat' structures where the information is factorised into minimal units. This facilitates processing and is key to the approach to underspecification.

- Elementary predicates correspond to morphological stems, annotated with 'v', 'n' etc to give a coarse-grained indication of sense.
- The POS tagged text shares the same lexicalised elementary predicates as the deep parser output (_mixture_n, _allow_v, _warm_v, _the_q), although the deep parser can make more fine-grained sense distinctions (allow_v_1) and inserts grammatical predications such as prpstn_m_rel (proposition).
- The POS tagger has no relational information (indicated in the deep output by ARG1 etc).
- The qeq conditions in the deep output are partial scope constraints which relate the 'h' labels.
- Uninstantiated relational positions in the deep output are indicated by 'u's, 'e's are eventualities, and 'x's other entities.
- For space reasons, we have shown the rendered form, rather than RMRS-XML, and have omitted much information including tense and number.

RMRS has been designed to be suitable for natural language representation and as such has to be very expressive while at the same time allowing for underspecification. Formally, RMRSs (like MRSs) are partial descriptions which correspond to a set of logical forms in a higher-order base language. RMRS itself is a restricted first order language: scope relationships are reified (via the 'h' labels) and natural language quantifiers, such as *every* and *most*, correspond to predicates, though these in turn correspond to generalised quantifiers in the base language.[1] Inference in the base language will not, in general, be tractable, but some inferences can be directly expressed using RMRS without resolving to the base language. RMRSs can be linked to ontologies, so that the notion of underspecification of an RMRS reflects the hierarchical ontological relationship. RMRS is thus distinct from RDF/OWL (`http://www.w3.org/TR/rdf-primer/`, `http://www.w3.org/TR/owl-features/`), but there are interesting formal correspondences. For the applications in this project, such as IE, RDF/OWL terms will be extracted from RMRSs.

RMRS has been developed on the EU project Deep Thought (`http://www.project-deepthought.net/`) over the last two years, primarily at Cambridge.[2] A range of deep and shallow processing systems have already been adapted to produce RMRS-XML and various tools exist for manipulating RMRSs. Initial results from Deep Thought partners demonstrate that RMRS can be successfully used for several types of IE, improving over the results available with the standard methods.

2.1 Development of RMRS

We will use the following shallow, intermediate and deep processors which the Cambridge group has already modified to produce RMRS. The cited processing speeds are approximate, based on a 1Ghz Pentium running Linux with 2 Gbyte of RAM:

- RASP part of speech tagger (Briscoe and Carroll, 2002): statistically determines tags for individual tokens in a text — 10,000 words/sec.
- RASP (Briscoe and Carroll, 2002): a statistically-trained parser which operates without a full lexicon — 100 words/sec.
- ERG (Copestake and Flickinger, 2000) processed by LKB(Copestake, 2002) or PET(Callmeier, 2002) deep processing which incorporates a lexicon with detailed linguistic information. PET is highly optimised (5–30 words/sec, depending on corpus) while the LKB is more suited for development and can be used for generation. These tools are all Open Source.

These systems and others were combined in Deep Thought via the Heart-of-Gold system developed at Saarbrücken (Callmeier et al., 2004) which we may adopt for this project. In the course of this project, we expect to require other tools, for NP-chunking, word sense disambiguation and anaphora resolution, for instance. We will either adapt existing Open Source software to RMRS or develop technology in-house on the basis of published algorithms.

The systems described are not domain-specific. While they have not yet been applied to chemistry texts, we will be able to take advantage of CML processing to isolate specialist terms.

One area where RMRS needs further development is in strategies for generalising over multiple outputs produced for ambiguous sentences. Both RASP and the ERG can produce thousands of readings and the accuracy of the stochastic techniques is not sufficiently high to make it a viable option to take only the highest

[1] In this respect, MRS/RMRS might be better regarded as a quasi-semantic representation (Alshawi and Crouch, 1992).

[2] Because Deep Thought was only a two year project, relatively little work has been published so far. Various working papers are accessible via `http://www.cl.cam.ac.uk/aac10/rmrs/`.

ranked analysis in all circumstances. The factorised nature of RMRS will allow us to use strategies that involve taking weighted intersections of structures: we expect the best strategy to depend on the needs of the application (in particular the balance between precision and recall), but this requires investigation (cf. (Carroll and Briscoe, 2002)).

We will adapt the tools which we use to construct RMRSs so that we can distribute processing over CamGrid and thus scale up to very large collections of papers. In the course of the project, we will be processing thousands of papers a week. Coarse-grained parallelism will be adequate, since we can distribute sections of text to different processors (see §2.6).

While RMRS could be used as the basis for many applications, we will concentrate here on three which all depend on matching patterns specified in terms of RMRS. The first is concerned with extraction of knowledge from texts to build a database of papers and key concepts fully automatically. The second is to use texts to semi-automatically construct ontologies of chemical concepts, expressed in OWL. The third application is geared towards humans browsing papers and to help them quickly see the most salient points and the interconnections between papers. These technologies, which are described in more detail in the next three sections, will be combined with others in the chemistry researchers' amanuensis, see §2.5.

2.2 Information extraction

Our most basic objective is to demonstrate that we can develop IE techniques to instantiate databases and knowledge bases with respect to a known ontology (cf., the GATE project (`http://gate.ac.uk`) 'Ontology Based Information Extraction/OBIE'). We will describe patterns in terms of RMRSs rather than on the basis of regular expressions over strings (see also (Surdeanu et al., 2003)). Patterns based on RMRS allow greater flexibility compared to text patterns, since one RMRS pattern can stand in for a large range of regular expressions over text. Pattern-matching is essentially based on semantic compatibility, rather than string matching. Consider for example the following two sentences, taken from an organic synthesis paper [3]:

> The reaction mixture was warmed to rt, whereat it was stirred overnight.
> The resultant mixture was kept at 0C for 0.5 h and then allowed to warm to rt over 1 h.

The RMRS pattern shown below would identify the fact that, in both cases, it is the mixture that is warmed (the ?s correspond to variables which are irrelevant to the pattern).

```
_mixture_n(?,x1),_warm_v(h2,?), ARG1(h2,x1)
```

Writing a finite state pattern that covered both sentences above while disallowing spurious matches cannot be done in an effective way, due to the control verb.

As we add further NL processing, the RMRSs that we can extract from text will be further refined. So, for instance, anaphora resolution would mean that some coreferences were resolved, allowing matching over an RMRS derived from separate sentences. Word sense disambiguation (WSD) will be an important technology: in particular we would expect to disambiguate words with respect to terms in ontologies. Note that, because we treat everything in terms of a refinement of RMRSs, addition of an anaphora resolution or WSD module should not require changes in the IE patterns.

Our first IE target will be defined in consultation with our partners, bearing in mind evaluation requirements, but eventually the amanuensis will not only support IE patterns defined by an NLP expert, but will allow users to create their own to address individual information needs.

2.3 Ontology construction

Because existing ontologies for Chemistry are limited, we will investigate semi-automatic ontology construction. As in §2.2, identification of ontological relationships will rely on patterns expressed in RMRS. Hearst (1992) developed an approach to taxonomy construction that exploited text cues. For example, 'is a' and 'and other' act as good cues:

... the concise synthesis of naturally occurring alkaloids and other complex polycyclic azacycles. gives 'alkaloid IS-A azacycle'.

But, as for the IE task, RMRS allows the specification of patterns that rely on identification of semantic concepts rather than directly encoding text cues which indirectly and ambiguously correspond to these concepts. For example, the following examples show 'is a' phrases which do not introduce taxonomic relationships.

... the combination of bis(pyridinium)ethane axles and 24-membered crown ethers is a versatile motif for forming [2]pseudorotaxanes, ...

[3] Bunnage, Mark, Stephen G. Davies, Paul M. Roberts, Andrew D. Smith, and Jonathan M. Withey. 2004. Asymmetric synthesis of the /cis/- and /trans/-stereoisomers of 4-aminopyrrolidine-3-carboxylic acid and 4-aminotetrahydrofuran-3-carboxylic acid. *Org. Biomol. Chem.* 2(19): 2763–2776.

Chemistry: The primary aims of the present study are (i) the synthesis of an amino acid derivative that can be incorporated into proteins /via/ standard solid-phase synthesis methods, and (ii) a test of the ability of the derivative to function as a photoswitch in a biological environment. Lougheed et al. (2004): 'Photomodulation of ionic current through hemithioindigo-modified gramicidin channels', *Org. Biomol. Chem*, Vol. 2, No. 19, 2798-2801
Computational Linguistics: The goal of the work reported here is to develop a method that can automatically refine the Hidden Markov Models to produce a more accurate language model. Kim et al. (1999): HMM Specialization with Selective Lexicalization, *EMNLP-99*

Figure 1: Similar phrases across the domains of chemistry and computational linguistics

. . . serine is a promising candidate and will be used in the next generation of modified analogues.

These are not IS-A because *motif* and *candidate* are not semantically 'kind' terms (in this use). We can generically specify that a kind is required in the RMRS pattern and develop techniques for distinguishing kind and non-kind expressions. Note that kind/non-kind distinction is relevant for other tasks, such as coreference resolution, so such markup is reusable.

2.4 Research markup

Searching in unfamiliar scientific literature is hard, even when a relevant paper is known as a starting point. One reason is that the status of a given paper with respect to similar papers is often not apparent from its abstract or the keywords it contains. For instance, a chemist might be more interested in papers containing direct experimental evidence rather than evidence by simulation, or might look for papers where some result is contradicted. Such subtle relationships between the core claims and evidence status of papers are currently not supported by search engines such as CiteSeer (`www.citeseer.com`); if we were able to model them, this would add considerable value.

The best sources for this information are the papers themselves. Discourse analysis can help, via an analysis of the argumentation structure of the paper. For instance, the author would typically follow the strategy of first pointing to gaps in the literature before describing the specific research goal – thereby adding important contrastive information in addition to the description of the research goal itself. An essential observation in this context is that conventional phrases are used to indicate the rhetorical status of different parts of a text. For instance, in Fig. 1 similar phrases are used to indicate the introduction of a goal, despite the fact that the papers come from different scientific domains.

Argumentative Zoning, a method introduced by Teufel, uses cues and other superficial markers to pick out important parts of scientific papers and supervised machine learning to find zones of different argumentative status in the paper. It was originally developed for computational linguistics (CL) papers, but as a general method of analysis, argumentative zoning can and has been applied to different text types (e.g., legal texts (Grover et al., 2003)), languages (e.g., Portuguese (Feltrim et al., 2005)) and text types (e.g., biological texts (Mizuta and Collier, 2004)); we will adapt it here to the special language of chemistry papers, and to the specific search tasks in eChemistry.

The zones used in the original (CL) domain concentrated on the phenomena of attribution of authorship to claims (is a given sentence an original claim of the author, or a statement of a well-known fact) and of citations sentiment (does the author criticise a certain reference or use it as part of their own work).[4] We will explore the utility of various between–paper relationships for eChemists' literary searches. Modelling these might require new types of zones. As concrete output, this research would lead to an automatic annotation of the chemistry literature with 'research markup', i.e., chemistry-specific annotation in terms of Argumentative Zones, attributed to each sentence. This will be exploited in the application we present in §2.5.

In terms of the features used, changes need to be made which mirror the different writing and argumentation styles in chemistry, in comparison to computational linguistics. Argumentation patterns are generally similar across the disciplines (there to convince the reader that the work undertaken is sound and grounded in evidence), but several factors such as the use of citations, passive voice, or cue phrases vary across domains.

This work will be closely coupled with the work on RMRS proposed earlier. As with IE, RMRS encoding is advantageous because it allows more concise and flexible specification of cues than with the string-based patterns and because it allows identification of more complex cases. Furthermore, deep processing of AZ cue phrases should be feasible because their vocabulary and structure is relatively consistent over a wide range of

[4]The following seven zones exist: AIM (the specific research goal of the current paper); TEXTUAL (statements about section structure); OWN (neutral descriptions of own work presented in current paper); BACKGROUND (generally accepted scientific background); CONTRAST (comparison with or contrast to other work); BASIS (statements of agreement with other work or continuation of other work); and OTHER (neutral descriptions of other researchers' work).

texts. In order to allow comparison with the existing AZ work, the port to RMRS will be done with cues from computational linguistics papers, and the chemistry cues will be directly encoded in RMRS.

We will investigate how the porting of features can be partially automated. Adapting the lexical features (such as the cue phrases) accounts for most of the porting work; discovering how new phrases could be automatically learned, rather than manually coded, is a new and challenging avenue of research.

This work will also profit from Teufel's EPSRC first grant "Rhetorical citation maps and domain-independent argumentative zoning", which will make the Argumentative Zoning software more robust towards variations in text type, build a large annotated corpus with argumentative zoning annotation, and explore argumentatively classified citation links.

2.5 The Chemists' Amanuensis

The modern e-science researcher publishes her results, her data (and in fact her entire working environment) in a re-usable way which is openly accessible to her fellow researchers. In turn, she can search for others' papers, data and results online, in order to directly use them for her own research, via the web or repositories such as DSpace (an interactive and searchable repository of papers and molecules, cf. `http://www.dspace.cam.ac.uk/`). In principle, large amounts of valuable peer-reviewed scientific information is at her disposal. We will build in this project a knowledge integration and search tool to demonstrate how this information could be exploited: the Chemists' Amanuensis.

The first version of the Amanuensis will be built on the basis of the Chemistry group's existing technology, with results from the NLP technology being incrementally incorporated. We will start off with simple IE and ontology extraction. The longer term goal of allowing the individual researcher to build their own IE patterns will eventually lead to individualised knowledge bases. The final version of the amanuensis will support continuously running IE processes, mining the corpus to identify information relevant to the user. The results may be updated as further information is acquired: for instance, adding a concept to an ontology which is subsumed by a node in a current IE pattern should trigger new search since additional matches are now possible.

The amanuensis is also a literature search tool, particularly for those search tasks where IE and research markup can be exploited. The ontologies will allow constrained inference which will support comparisons to find similarities and contradictions in passages identified by the research markup. Since research markup is a novel concept, the determination of exactly how it will be exploited will involve experimentation and discussion with potential users, but here are some indicative tasks:

Which researchers form a subcommunity?

Researchers tend to form communities with members who know and cite each other's work. Since research markup makes clear the context of citations, these relationships can be discovered by researchers from outside the field, editors seeking reviewers, etc.

Find papers like this one, but which differ in aspect X.

Research markup can provide an operational model of 'similarity' above and beyond simple string/keyword similarity, by isolating passages describing goals and methods. A fine-grained notion of similarity can also serve to identify genuinely novel papers: they have goals which are maximally different from those of established papers.

Which papers in the repository present contradictory evidence to my own results?

On the basis of research markup of results and evidence status, contradictions could be identified in passages expressing results based on simple models of negation, ontological knowledge about antonymy and concepts which exclude each other.

Although the bulk of this proposal has been concerned with published text, we hope to experiment in the amanuensis with other types of science-related text, including safety regulations, grant proposals, reviews and patents.

2.6 Applying Grid computing

Processing the corpora of texts provided by the publishers will require the use of high throughput Grid computing. The Cambridge eScience Centre will provide the necessary resources to do this in the form of the CamGrid infrastructure (a University-wide Grid based on Condor) and the workflow tools required to go from notification of publication (e.g., a RSS feed provided by Nature) to the results required for the Chemists' Amanuensis application.

CamGrid consists of computer clusters in research groups from across the University. The Condor batch system allows users to submit jobs to the most appropriate resource that is currently free. The University Computing Service are also deploying Condor on teaching workstations across the University. This represents a significant computational resource suitable for coarse-grained, pleasingly parallel tasks such as natural language

processing. CeSC will provide on-going support throughout the project to maintain CamGrid as a production environment.

There is a clear requirement to automate the processing of the initial corpus and new publications as they appear. Workflow tools such as Condor's DAGMAN can be used to co-ordinate the transfer of data and the submission of the associated computational task. Integration with external sources of data such as RSS notifications from the publishers will also be required. CeSC will support the development and use of these tools, providing the necessary infrastructure for the domain-specific work at both departments.

3 Relevance

This project has both short-term and long-term goals. The short-term goal is the development of the Chemist's Amanuensis tool, which will be useful to anyone dealing with the published Chemistry literature. We expect its browsing aspects to be particularly beneficial to researchers working in isolated groups, who cannot easily discuss the research field with more experienced colleagues. It will thus enhance the ability of people in SMEs and small university departments to connect to the research community. It will improve the way that researchers in all levels of organisation exploit the existing literature. Besides Chemistry researchers, the Amanuensis will benefit publishers, government organisations and researchers in other disciplines who have to navigate Chemistry texts. Successful development would lead to similar approaches being possible for other scientific disciplines.

Our long-term challenge is the development of a standardised approach to markup which can be produced by a wide-range of natural language processing tools and act as a common interchange language. We hope to use this project as a basis to extend our current network of collaborators in the UK and overseas and to develop a community of researchers who will share and build on the methodology. Distribution of the Open Source technology developed on the project will benefit other researchers in NLP. By providing a truly integrated set of tools, it will become much easier to build sample applications and test novel methods. In the long term, this could have significant implications for NLP technology because it would make it much more accessible to people outside the research community. NLP technology is currently very limited in its applications by the specialist nature of the skills required to deploy it. If we can reduce that overhead, a much wider range of applications will be attempted. Ultimately this would benefit anyone who needed to extract information from the web. Of course, we do not expect to achieve this ambitious objective by work on this project alone, but we will only convince people to adopt a common approach by demonstrating its worth in a truly large-scale practical application, for which eScience is ideal.

The Open Source nature of the technology is essential for the collaborative enterprise, which in turn is required for success, since in the long-term the NLP work is beyond the scale of anything that might be accomplished by a single group. However, by basing this project in the UK, we will develop a pool of UK expertise which will facilitate access to the research by UK businesses who wish to exploit the technology to build commercial applications.

References

Alshawi, H., and R. Crouch. 1992. Monotonic Semantic Interpretation. In *Proc. ACL-92*.

Briscoe, E.J., and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of LREC-2002*.

Callmeier, U.. 2002. Pre-processing and encoding techniques in PET. In S. Oepen, D. Flickinger, J. Tsujii, and H. Uszkoreit, eds., *Collaborative Language Engineering: a case study in efficient grammar-based processing*. Stanford: CSLI Publications.

Callmeier, U., A. Eisele, U. Schäfer, and M. Siegel. 2004. The DeepThought Core Architecture Framework. In *Proc. of LREC-2004*.

Carroll, J., and E.J. Briscoe. 2002. High Precision Extraction of Grammatical Relations. In H. Bunt, J. Carroll, and G. Satta, eds., *Proc. of Int. Workshop on Parsing Technologies*. Kluwer.

Copestake, A. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Copestake, A. 2003. Report on the design of RMRS. DeepThought project deliverable.

Copestake, A., and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proc. of LREC-2000*, 591–600.

Copestake, A., D. Flickinger, R. Malouf, S. Riehemann, and I. Sag. 1995. Translation using Minimal Recursion Semantics. In *The Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*.

Copestake, A., D. Flickinger, I. Sag, and C. Pollard. In Press. Minimal Recursion Semantics: An introduction. *Research in Language and Computation*.

Copestake, A., A. Lascarides, and D. Flickinger. 2001. An Algebra for Semantic Construction in Constraint-based Grammars. In *Proc. of ACL-01*.

Feltrim, V., S. Teufel, G. G. Nunes, and S. Alusio. 2005. Argumentative Zoning applied to Critiquing Novices' Scientific Abstracts. In James G. Shanahan, Yan Qu, and Janyce Wiebe, eds., *Computing Attitude and Affect in Text*. Dordrecht, The Netherlands: Springer.

Grover, C., B. Hachey, and C. Korycinsky. 2003. Summarising legal texts: Sentential tense and argumentative roles. In *Proc. of the NAACL/HLT-03 Workshop on Automatic Summarization*.

Hearst, M. A. 1992. Direction-Based Text Interpretation as an Information Access Refinement. In P. S. Jacobs, ed., *Text-based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. Hillsdale, NJ: Lawrence Erlbaum.

Mizuta, Y., and N. Collier. 2004. An Annotation Scheme for Rhetorical Analysis of Biology Articles. In *Proc. of LREC'2004*.

Murray-Rust, P., R. C. Glen, H. S. Rzepa, J. J. P. Stewart, J. A. Townsend, E. L. Willighagen, and Y. Zhang. 2003. A semantic GRID for molecular science. In *Proc. of UK e-Science All Hands Meeting*, 802–809.

Murray-Rust, P., and H. S. Rzepa. 2003. Towards the Chemical Semantic Web. An introduction to RSS. *Internet J. Chem.* 6.

Murray-Rust, P., and H. S. Rzepa. 2004. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information* 5.

Murray-Rust, P., H. S. Rzepa, M. J. Williamson, and E. L. Willighagen. 2004a. Chemical Markup, XML and the Worldwide Web. Part 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comp. Sci.* 44: 462–469.

Murray-Rust, P., H. S Rzepa, S. M. Tyrrell, and Y. Zhang. 2004c. Representation and use of Chemistry in the Global Electronic Age. *Org. Biomol. Chem.*

Surdeanu, M., S. M. Harabagiu, J. Williams, and P. Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proc. of ACL-03*, 8–15.

Teufel, S., J. Carletta, and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proc. of EACL-99*, 110–117.

Teufel, S., V. Hatzivassiloglou, K. McKeown, K. Dunn, D. Jordan, S. Sigelman, and A. Kushniruk. 2001. Personalized Medical Article Selection using Patient Record Information. In *Proc. of AMIA-2001*.

Teufel, S., and M. Moens. 2000. What's yours and what's mine: Determining Intellectual Attribution in Scientific Text. In *Proc. of EMNLP-00*.

Teufel, S., and M. Moens. 2002. Summarising Scientific Articles — Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 28(4): 409–446.

Townsend, J. A., S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust. 2004. Chemical documents: machine understanding and automated information extraction. *Org. Biomol. Chem.*