

SciBorg: Deep Processing and Chemical Informatics

Ann Copestake, Peter Corbett, CJ Rupp,
Advaith Siddharthan, Simone Teufel, Ben
Waldron

University of Cambridge

Overview

- semantic markup language for integrated processing
- introduction to the SciBorg project
- overview of architecture
- semantic markup in SciBorg
- domain-dependent modules
- citation classification
- conclusion

Compositional semantics as a common representation for NLP integration

- Different NLP systems have different strengths and weaknesses
- Pairwise compatibility between systems is too limiting
 - Syntax is theory-specific and too language-specific
 - Eventual goal should be semantics
- Core idea: shallow processing gives underspecified semantic representation with respect to a normative `deep' analysis
- Integrate processors with different capabilities
- Applications work on a standard representation
- Reuse of knowledge sources, integration with ontologies
- First experiments done on Deep Thought and QUETAL: RMRS language

Extracting the science from scientific publications: SciBorg

- 4-year EPSRC-funded project started in October 2005
 - Computer Laboratory, Chemistry, Cambridge eScience Centre
 - Nature Publishing, Royal Society of Chemistry, International Union of Crystallography (papers and publishing expertise)
- Aims:
 1. Develop an NL markup language (RMRS) which will act as a platform for extraction of information. Link to semantic web languages.
 2. Develop IE technology and core ontologies for use by publishers, researchers, readers, vendors and regulatory organisations.
 3. Model scientific argumentation and citation purpose in order to support novel modes of information access.
 4. Demonstrate the applicability of this infrastructure in a real-world eScience environment.

General assumptions

- There is lots of useful information in the published scientific literature that is not currently being retrieved
- Language processing is required for some sorts of analyses (text-mining versus data-mining)
- Building specialized language processing tools for each task isn't cost-effective (time and skill), so we need to build and exploit general purpose language technology
- Eventually language technology should be a standard part of Computer Science, like database technology: i.e., needs some time and expertise to adapt to new tasks and domains, but not (as currently) a research project
- Text processing tools based directly on text patterns (regular expressions) work adequately for some tasks, but often fail to achieve high enough precision and recall

Variation in expression

Example 1: **searching for papers describing synthesis of Tröger's base from anilines:**

A: The synthesis of 2,8-dimethyl-6H,12H-5,11-methanodibenzo[b,f][1,5]diazocine (Troger's base) from p-toluidine and of two Troger's base analogs from other anilines

B: ... Tröger's base (TB) ... The TBs are usually prepared from para-substituted anilines

linguistic variation and syntactic relationship (**synthesis of X, synthesize X, prepare X and so on**), coreference, chemistry names, ontological information ...

Example 2: **searching for papers describing Tröger's base syntheses which don't involve anilines.**

SciBorg, or the Chemist's amanuensis

- Research prototype, bringing together different language processing tools supporting different types of information extraction (IE)
- Process chemistry texts using combined domain-independent and domain-dependent processing: markup in RMRS
- IE based on patterns expressed via semantics and rhetorical organization:

retrieve all papers X: PAPER-AIM(X,h), h:synthesis,
SYN-RESULT(h,<TB>), SYN-SOURCE(h,y) &
NOT(aniline(y))

Information Extraction

Chemistry IE: e.g., Organic chemistry syntheses

To a solution of aldimine¹ (1.5mmol) in THF (5mL) was added LDA (1mL, 1.6 M in THF) at 0 °C under argon, the resulting mixture was stirred for 2h, then was cooled to -78 °C ...

➡ recipe expressed in chemistry formalism (CML)

Ontology extraction (to support other IE)

... alkaloids and other complex polycyclic azacycles ...

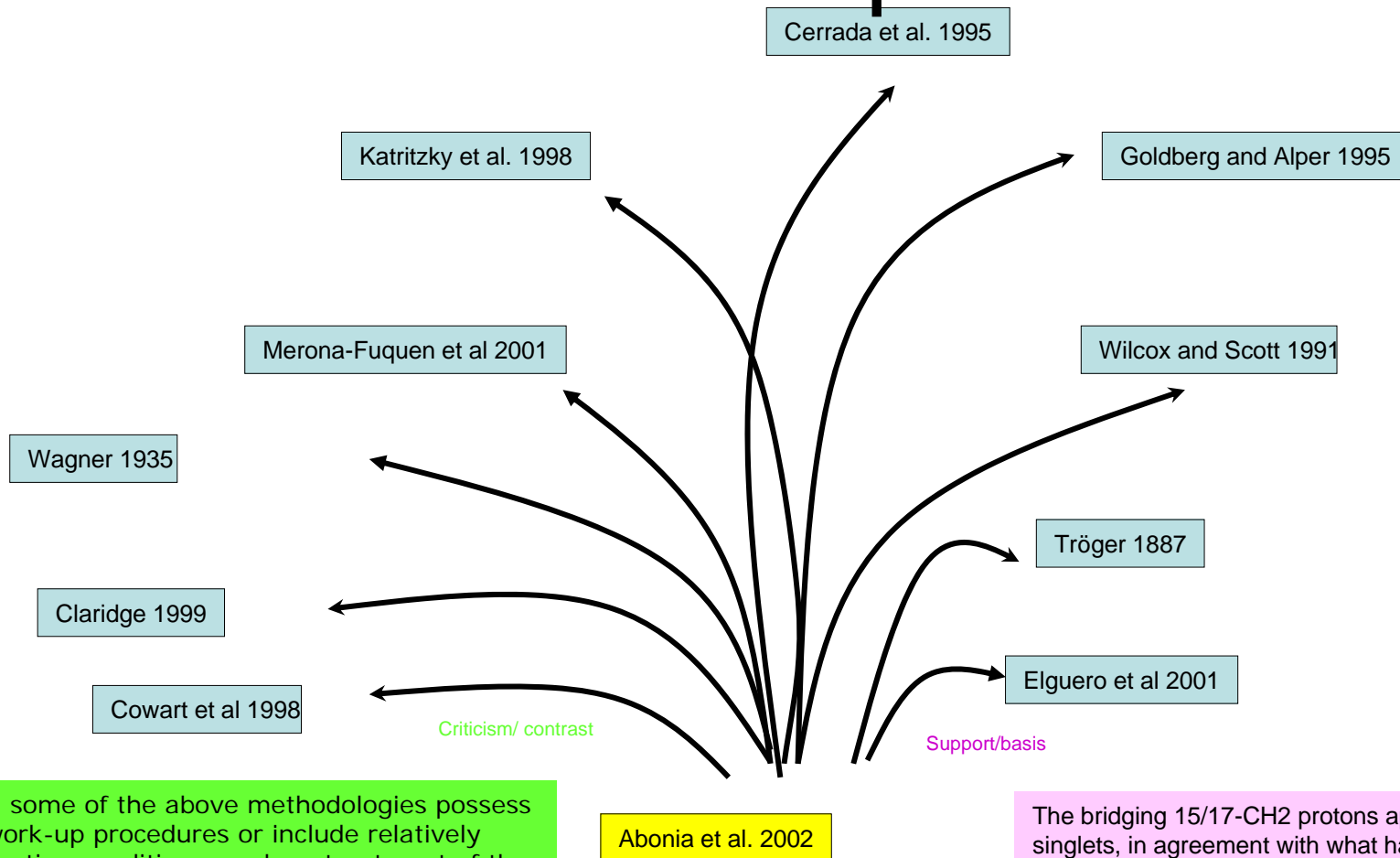
➡ `<owl:Class rdf:ID="Alkaloid">`
`<rdfs:subClassOf rdf:resource="#Azacycle" />`

Research markup

Enamines have been used widely ... (citation Y), however, ... did not provide the desired products.

➡ X cites Y (contrast)

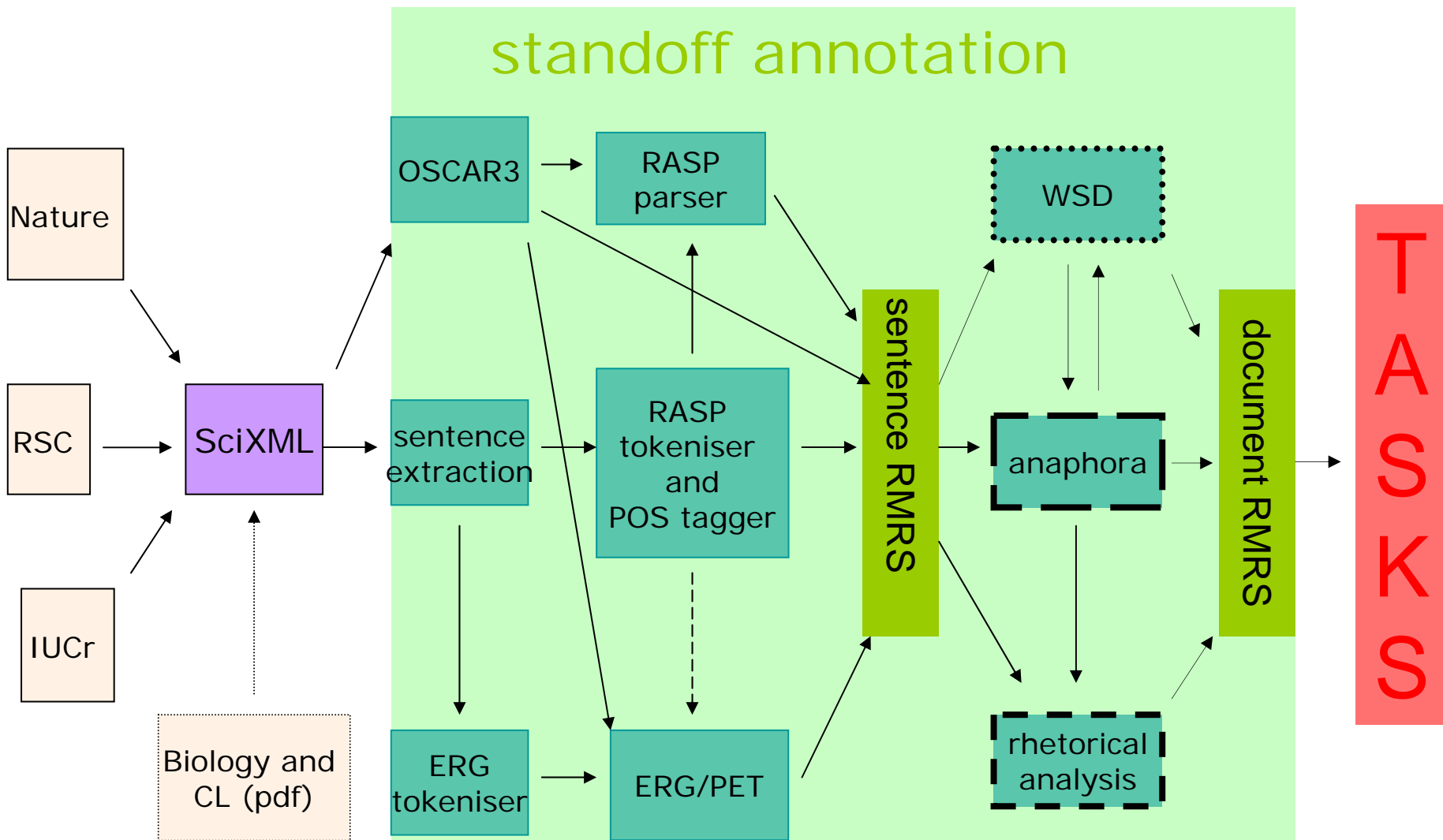
Citation map



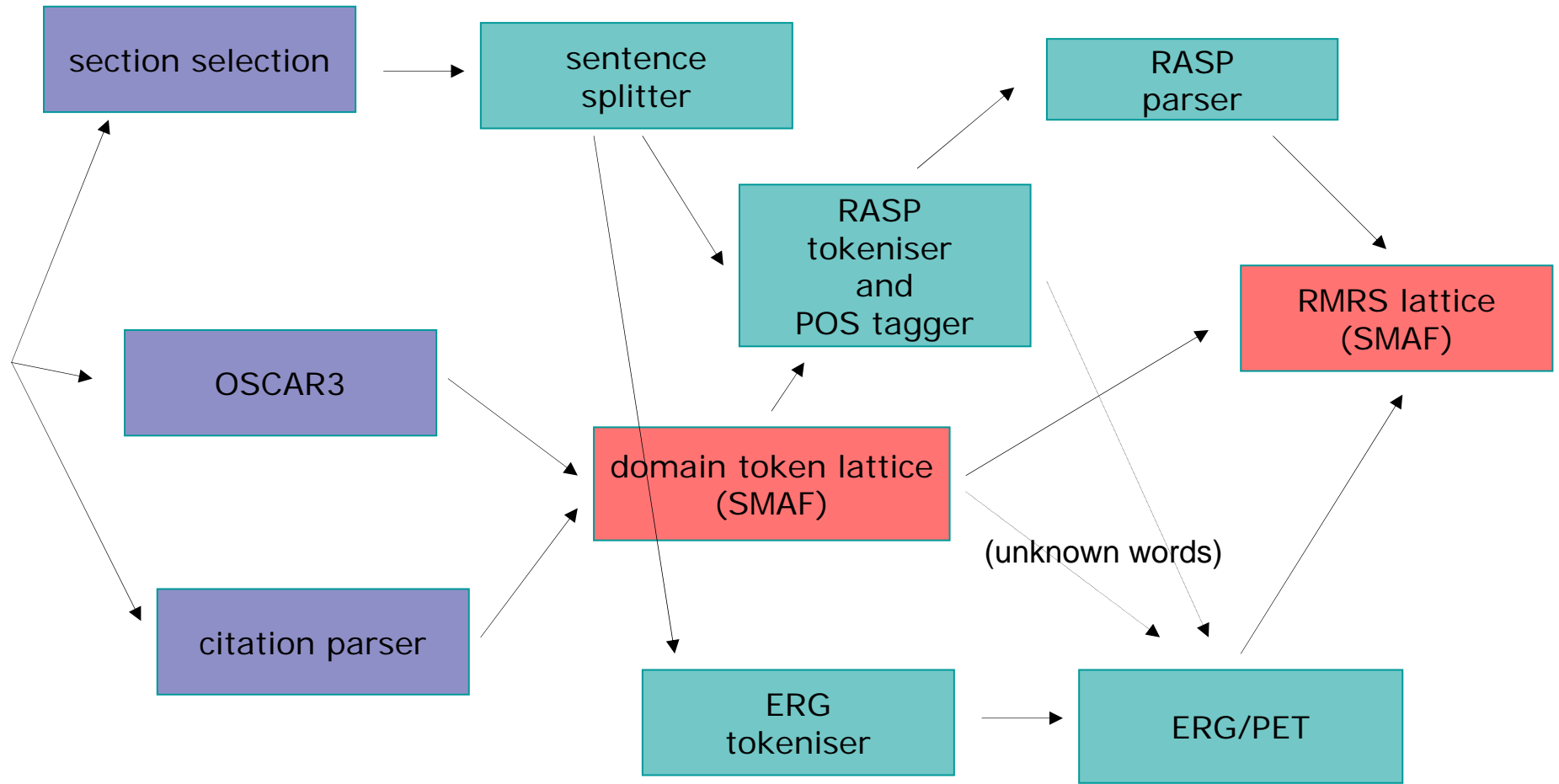
However, some of the above methodologies possess tedious work-up procedures or include relatively strong reaction conditions, such as treatment of the starting materials for several hours with an ethanolic solution of conc. hydrochloric acid or TFA solution, with poor to moderate yields, as is the case for analogues **4** and **5**.

The bridging 15/17-CH₂ protons appear as singlets, in agreement with what has been observed for similar systems [9].

Outline architecture



Details of sentence parsing



SciXML: text markup for scientific papers

```
<?xml version="1.0" encoding="UTF-8"?>
<PAPER>
  <METADATA> <FILENO>b200862a</FILENO>
    <JOURNAL><NAME>P1</NAME><YEAR>2002</YEAR>
    <ISSUE>13</ISSUE> <PAGES>1588-1591</PAGES></JOURNAL>
  </METADATA>
  <TITLE>Synthesis of pyrazole and pyrimidine Tröger's-base analogues</TITLE>
  <AUTHORLIST><AUTHOR
    ID="1">Rodrigo<SURNAME>Abonia</SURNAME></AUTHOR>           <AUTHOR
    ID="2">Andrea<SURNAME>Albornoz</SURNAME></AUTHOR>...
  </AUTHORLIST>
  <ABSTRACT>Tröger's-base analogues bearing fused pyrazolic or pyrimidinic rings
  were prepared in acceptable to good yields through the reaction of 3-alkyl-5-amino-1-
  arylpyrazoles and 6-aminopyrimidin-4(3<IT>H</IT>)-ones with formaldehyde under
  mild conditions (<IT>i.e.</IT>, in ethanol at 50 ° C in the presence of catalytic
  amounts of acetic acid). Two key intermediates were isolated from the reaction
  mixtures, which helped us to suggest a sequence of steps for the formation of the
  Tröger's bases obtained. The structures of the products were assigned by
  <SP>1</SP> H and <SP>13</SP>C NMR, mass spectra and elemental analysis
  and confirmed by X-ray diffraction for one of the obtained compounds.</ABSTRACT>
```

Domain-independent language processing

- ERG (English Resource Grammar)/PET
 - DELPH-IN technology (www.delph-in.net), Open Source
 - LKB for grammar development (and generation), PET for fast parsing
 - HPSG, stochastic ranking
 - detailed lexicon, various approaches to unknown words
 - max coverage about 80% on general text, tuning required for some constructions, relatively slow (100 words/sec)
 - Minimal Recursion Semantics (MRS) output, converted to RMRS
- RASP 2
 - Briscoe and Carroll et al
 - initial POS tagging stage, symbolic grammar over tags (hand-written), stochastic ranking, no lexicon required
 - robust to missing lexical entries, faster (1000 words/sec), relatively shallow
 - RASP-RMRS (Deep Thought/SciBorg DELPH-IN licence)

Simplified RMRS example:

'the mixture was allowed to warm'

- ERG-RMRS

_the_q (h1,x2)
RSTR(h1,h3)
BODY(h1,h8)
_mixture_n(h3,x4)
ARG1(h3,u10)
_allow_v_1(h5,e6)
ARG1(h5,u11)
ARG2(h5,x3)
ARG3(h5,h8)
qeq(h8,h7)
_warm_v(h7,e8)
ARG1(h7,x4)
x2=x4

- RASP-RMRS

_the_q (h1,x2)
RSTR(h1,h3)
BODY(h1,h8)
_mixture_n(h3,x4)
_allow_v(h5,e6)
ARG2(h5,x3)
ARG3(h5,h8)
qeq(h8,h7)
_warm_v(h7,e8)
x2=x4

- POS-RMRS

_the_q (h1,x2)
_mixture_n(h3,x4)
_allow_v (h5,e6)
_warm_v(h7,e8)

More RMRS detail

- ERG (R)MRS can undergo scope resolution to produce conventional logical forms
- other forms of specialisation: WSD, compound relation identification etc
- arguments (not just for verbs), complex quantifiers, prepositions (vs particles), construction semantics etc
- RMRS allows underspecification of ARG (e.g., ARG2-3): useful for `deep' grammars too
- standoff annotation allows RMRSs from different sources to be aligned (sloppy alignment)

<ep cfrom='0' cto='4'><realpred lemma='some' pos='q' /><label vid='3' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='0' cto='4'><gpred>part_of_rel</gpred><label vid='7' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='8' cto='11'><realpred lemma='the' pos='q' /><label vid='9' />
<var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='12' cto='26'><gpred>compound_rel</gpred><label vid='12' />
<var sort='e' vid='14' tense='u' /></ep>
<ep cfrom='12' cto='26'><gpred>undef_q_rel</gpred><label vid='15' />
<var sort='x' vid='13' /></ep>
<ep cfrom='12' cto='17'><realpred lemma='train' pos='n' sense='of' />
<label vid='18' /><var sort='x' vid='13' /></ep>
<ep cfrom='18' cto='26'><realpred lemma='station' pos='n' sense='1' />
<label vid='10001' /><var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='27' cto='33'><gpred>neg_rel</gpred><label vid='20' />
<var sort='e' vid='22' tense='u' /></ep>
<ep cfrom='39' cto='46'><realpred lemma='check' pos='v' sense='1' />
<label vid='23' /><var sort='e' vid='2' tense='past' /></ep>
<ep cfrom='47' cto='55'><gpred>unspec_loc_rel</gpred><label vid='10002' />
<var sort='e' vid='26' tense='u' /></ep>
<ep cfrom='47' cto='55'><gpred>proper_q_rel</gpred><label vid='27' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>
<ep cfrom='47' cto='55'><gpred>dofw_rel</gpred><label vid='30' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>

<ep cfrom='0' cto='4'><realpred lemma='some' pos='q' /><label vid='3' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='0' cto='4'><gpred>part_of_rel</gpred><label vid='7' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='8' cto='11'><realpred lemma='the' pos='q' /><label vid='9' />
<var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='12' cto='26'><gpred>compound_rel</gpred><label vid='12' />
<var sort='e' vid='14' tense='u' /></ep>
<ep cfrom='12' cto='26'><gpred>undef_q_rel</gpred><label vid='15' />
<var sort='x' vid='13' /></ep>
<ep cfrom='12' cto='17'><realpred lemma='train' pos='n' sense='of' />
<label vid='18' /><var sort='x' vid='13' /></ep>
<ep cfrom='18' cto='26'><realpred lemma='station' pos='n' sense='1' />
<label vid='10001' /><var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='27' cto='33'><gpred>neg_rel</gpred><label vid='20' />
<var sort='e' vid='22' tense='u' /></ep>
<ep cfrom='39' cto='46'><realpred lemma='check' pos='v' sense='1' />
<label vid='23' /><var sort='e' vid='2' tense='past' /></ep>
<ep cfrom='47' cto='55'><gpred>unspec_loc_rel</gpred><label vid='10002' />
<var sort='e' vid='26' tense='u' /></ep>
<ep cfrom='47' cto='55'><gpred>proper_q_rel</gpred><label vid='27' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>
<ep cfrom='47' cto='55'><gpred>dofw_rel</gpred><label vid='30' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>

<ep cfrom='0' cto='4'><realpred lemma='some' pos='q' /><label vid='3' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='0' cto='4'><gpred>part_of_rel</gpred><label vid='7' />
<var sort='x' vid='4' pers='3' num='pl' /></ep>
<ep cfrom='8' cto='11'><realpred lemma='the' pos='q' /><label vid='9' />
<var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='12' cto='26'><gpred>compound_rel</gpred><label vid='12' />
<var sort='e' vid='14' tense='u' /></ep>
<ep cfrom='12' cto='26'><gpred>undef_q_rel</gpred><label vid='15' />
<var sort='x' vid='13' /></ep>
<ep cfrom='12' cto='17'><realpred lemma='train' pos='n' sense='of' />
<label vid='18' /><var sort='x' vid='13' /></ep>
<ep cfrom='18' cto='26'><realpred lemma='station' pos='n' sense='1' />
<label vid='10001' /><var sort='x' vid='8' pers='3' num='pl' /></ep>
<ep cfrom='27' cto='33'><gpred>neg_rel</gpred><label vid='20' />
<var sort='e' vid='22' tense='u' /></ep>
<ep cfrom='39' cto='46'><realpred lemma='check' pos='v' sense='1' />
<label vid='23' /><var sort='e' vid='2' tense='past' /></ep>
<ep cfrom='47' cto='55'><gpred>unspec_loc_rel</gpred><label vid='10002' />
<var sort='e' vid='26' tense='u' /></ep>
<ep cfrom='47' cto='55'><gpred>proper_q_rel</gpred><label vid='27' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>
<ep cfrom='47' cto='55'><gpred>dofw_rel</gpred><label vid='30' />
<var sort='x' vid='25' pers='3' num='sg' /></ep>

RMRS construction

- OSCAR-3: different types of chemical compound reference mapped to simple RMRSs (analogous to nouns etc)
- POS-RMRS: tag lexicon
- RASP-RMRS: tag lexicon plus semantic rules associated with RASP rules
 - no lexical subcategorization, so rely on grammar rules to provide the ARGs
 - developed on basis of ERG semantic test suite
 - default composition principles when no rule RMRS specified
- ERG-RMRS: converted from MRS
- Research Markup: RMRS versions of cue phrases

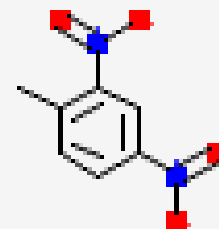
Chemistry naming

2,4-dinitrotoluene

Trivial name: (toluene), plus additional groups (dinitro) and positions (2,4)

Alternative names:

1-methyl-2,4-dinitro-benzene,
2,4-dinitromethylbenzene,
2,4-DNT and so on



toluene

Generic references: dinitrotoluenes

Chemistry Markup Language (CML, Murray-Rust et al)

- Language for formal, precise specification of organic chemistry structures in XML
- Language being actively extended
- Markup of chemistry papers with CML
- Already extensive online appendices to chemistry papers (spectra etc)
- Authoring tools for checking papers (e.g., checking that name used matches with spectrum)
- OSCAR-3: identification of productive chemistry terms and conversion to CML
- OSCAR-3: now in use by RSC journal publications

Oscar Annotations

- We use Oscar3 to identify possible chemical terms (and formatted data sections)
- Interpretations:
 - {compound, element, substance} -> nominal lexical entry (possibly plural)
 - reaction (e.g., *methylate*) -> verb (or nominalisation)
- Ambiguity: e.g., *lead*, *In*
- High recall, low precision mode: treat as token and sense ambiguity for ERG (and RASP?)

Research Markup for e-chemistry

- Better, rhetorically oriented search
 - “Find me contradictory claims to the ones in that paper”
- Improve automatic indexing (eg. CiteSeer)
 - At-a-glance map shows type of rhetorical relations between papers
 - Automatic classification rather than human perusing of each citation context
 - Which citations are more important in the paper?
 - What is the authors’ stance towards them?
 - Find “schools of thought”
- Difference and similarity-oriented summaries

Synthesis of pyrazole and pyrimidine Tröger's base analogues

Rodrigo Abonia, Andrea Alborno, Hector Larrahondo, Jairo Quiroga, Braulio Insuasty, Henry Insuasty, Angelina Hormaza, Adolfo Sánchez, Manuel Noguera

Tröger's-base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3-alkyl-5-amino-1-arylpyrazoles and 6-aminopyrimidin-4(3*H*)-ones with formaldehyde under mild conditions (*i.e.*, in ethanol at 50 °C in the presence of catalytic amounts of acetic acid).

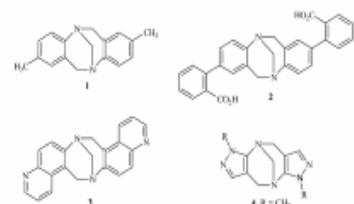
Two key intermediates were isolated from the reaction mixtures, which helped us to suggest a sequence of steps for the formation of the Tröger's bases obtained. The structures of the products were assigned by ¹H and ¹³C NMR, mass spectra and elemental analysis and confirmed by X-ray diffraction for one of the obtained compounds.

1
Perkin

Introduction

Although the first Tröger's base **1** was obtained more than a century ago from the reaction of *p*-toluidine and formaldehyde, [1] recently the study of these compounds has gained importance due to their potential applications. They possess a relatively rigid chiral structure which makes them suitable for the development of possible synthetic enzyme and artificial receptor systems, [2] chelating and biomimetic systems, [3] and transition metal complexes for regio- and stereoselective catalytic reactions. [4]

For these reasons, numerous Tröger's-base derivatives have been prepared bearing different types of substituents and structures (*i.e.*, **2–5** Scheme 1), with the purpose of



Scheme 1 The original Tröger's base **1** and some interesting derivatives and analogues.

However, some of the above methodologies possess tedious work-up procedures or include relatively strong reaction conditions, such as treatment of the starting materials for several hours with an ethanolic solution of conc. hydrochloric acid or TFA solution, with poor to moderate yields, as is the case for analogues **4** and **5**.

Considering these potential applications, we now report a simple synthetic method for the preparation of 5,12-dialkyl-3,10-diaryl-1,3,4,8,10,11-hexaazatetracyclo[6.6.1.0.2,6.0.9,13]pentadeca-2(6),4,9(13),11-tetraenes **8a–e** and 4,12-dimethoxy-1,3,5,9,11,13-hexaazatetracyclo[7.7.1.0.2,7.0.10,15]heptadeca-2(7),3,10(15),11-tetraene-6,14-diones **10a,b** based on the reaction of 3-alkyl-5-amino-1-arylpyrazoles **6** and 6-aminopyrimidin-4(3*H*)-ones **9** with formaldehyde in ethanol and catalytic

amounts of acetic acid. Compounds **8** and **10** are new Tröger's-base analogues bearing heterocyclic rings instead of the usual phenyl rings in their aromatic parts.

Results and discussion

In an attempt to prepare the benzotriazolyl derivative **7a**, which could be used as an intermediate in the synthesis of new hydroquinoline analogues of interest, [6] a mixture of 5-amino-3-methyl-1-phenylpyrazole **6a**, formaldehyde and benzotriazole in 10 mL of ethanol, with catalytic amounts of acetic acid, was heated at 50 °C for 5 minutes. A solid precipitated from the solution while it was still hot. However, no consumption of benzotriazole was observed by TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazole, as shown in Chart 1. On the basis of NMR and mass spectra and X-ray crystallographic analysis we established that the structure of this compound is 5,12-dimethyl-3,10-diphenyl-1,3,4,8,10,11-hexaazatetracyclo[6.6.1.0.2,6.0.9,13]pentadeca-2(6),4,9(13),11-tetraene **8a**, a new pentagonal Tröger's-

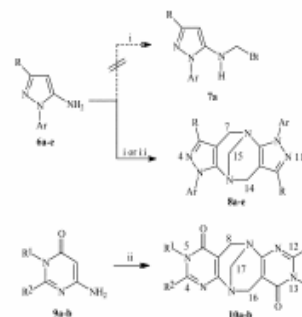


Chart 1 Reaction of 5-amino-3-methyl-1-phenylpyrazoles **6** and 6-aminopyrimidin-4(3*H*)-ones **9** with formaldehyde. Reagents: I = CH₂O, EtOH, EtOH; II = CH₂O, EtOH, HOAc; BH = Benzotriazole.

Legenda:

Background

Other

Own

Based

Contrast

Textual

Aim

1588 *J.Chem. Soc., Perkin Trans. 1, 2002, 1588-1591*

This journal is © The Royal Society of Chemistry 2002

DOI: 10.1039/b200862a

Research markup

- 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.
- 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.

RMRS and research markup

- Specify cues in RMRS: e.g.,
 - I1:**objective**(x), ARG1(I1,y), I2:**research**(y)
 - The concept **objective** generalises the predicates for *aim*, *goal* etc and **research** generalises *study*, *work* etc. Ontology for rhetorical structure.
- Deep process possible cue phrases to get RMRSs:
 - feasible because domain-independent
 - more general and reliable than shallow techniques
 - allows for complex interrelationships e.g.,
our goal is not to ... but to ...
- Use zones for advanced citation maps (e.g., X cites Y (contrast)) and other enhancements to repositories

Conclusion: extending technology in several ways

- SciXML (and standoff)
 - general framework for scientific texts
- more extensive and more varied IE-like operations
 - support for scientific discourse processing
 - ontology extraction
- finer-grained deep-shallow integration
 - deep cue phrase analysis
- unusual NER-like processing for chemistry with OSCAR3
- discourse level processing with DELPH-IN technology
 - anaphora, WSD, citations and research markup

Status of SciBorg aims

1. **NL markup language (RMRS).** Basic architecture for text processing in place (SciXML, standoff, lattices, OSCAR-3, RASP2 and ERG/PET). Next steps:
 - debugging scripts, regression test sets
 - Treebank with ERG (maybe use for evaluating RASP ranking too?)
 - RMRS lattices from packed representations?
 - use of CamGrid (coarse-grained parallelism)
2. **IE technology and core ontologies.** OSCAR-3 in use by RSC.
 - Initial experiments with ontology extraction based on RASP-RMRS from Wikipedia (Aurelie Herbelot).
3. **Model scientific argumentation and citation purpose.** Finding rhetorical cues with aid of RMRS (so far in CL papers only).
4. **Applicability in a real-world eScience environment.**
 - Partial change in emphasis to using technology for authoring support, based on publishers' interests.

Using external ontologies

- concepts like **research** generalizing *study, work* etc: automatic acquisition? (machine learning or FrameNet)
- IE is ontologically driven (some ontologies exist for Chemistry, but not as rich as biology, hence the need to augment)
- chemical naming provides implicit ontology
- ontologies bootstrapping ontology acquisition
- CML target for IE tasks
- classification of trivial chemistry names etc