

# Extracting the science from scientific publications

---

Simone Teufel

Computer Laboratory,

University of Cambridge

[sht25@cl.cam.ac.uk](mailto:sht25@cl.cam.ac.uk)

# Acknowledgements

---

- SciBorg (EPSRC, 2005-2009)
  - Computer Lab: Ann Copestake, Simone Teufel, CJ Rupp, Advait Siddharthan
  - Chemistry: Peter Murray-Rust, Peter Corbett
  - CeSC: Mark Hayes, Andy Parker
- CITRAZ (EPSRC, 2003-2006)
  - Computer Lab: Simone Teufel, Dan Tidhar, Anna Ritchie, Bill Hollingsworth
- FLYSLIP (BBSRC, 2004-2008)
  - Computer Lab: Ted Briscoe, Simone Teufel, Ian Lewin, Nikiforos Karamanis
  - Genetics: Rachel Drysdale plus FLYBASE curator team
- DELPH-IN (informal ongoing collaboration)
  - Boeing funding to Computer Lab: Ben Waldron
  - especially Dan Flickinger, Alex Lascarides, Stephan Oepen, John Carroll, Anette Frank
  - Prior work on Deep Thought (EU funded)

# Outline

---

- Objectives of SciBorg
- Overview of architecture
  - SciXML
  - SMAF
- Main components
  - OSCAR
  - RASP and PET/ERG
  - RMRS for integration
  - Links to ontologies, WSD, anaphora resolution (future work)
  - Research markup (AZ) - project CitRAZ
  - Citation analysis

# SciBorg: Chemistry texts

---

- eScience project started in October at Cambridge
  - Computer Laboratory, Chemistry, CeSC
  - Partners: Nature Publishing, Royal Society of Chemistry, International Union of Crystallography (supplying papers and publishing expertise)
- Aims:
  1. Develop an NL markup language 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.

# Information extraction in SciBorg

## Chemistry IE: e.g., Organic chemistry syntheses

To a solution of aldimine<sub>1</sub> (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 (extension of CML)

## Ontology extraction

... 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.

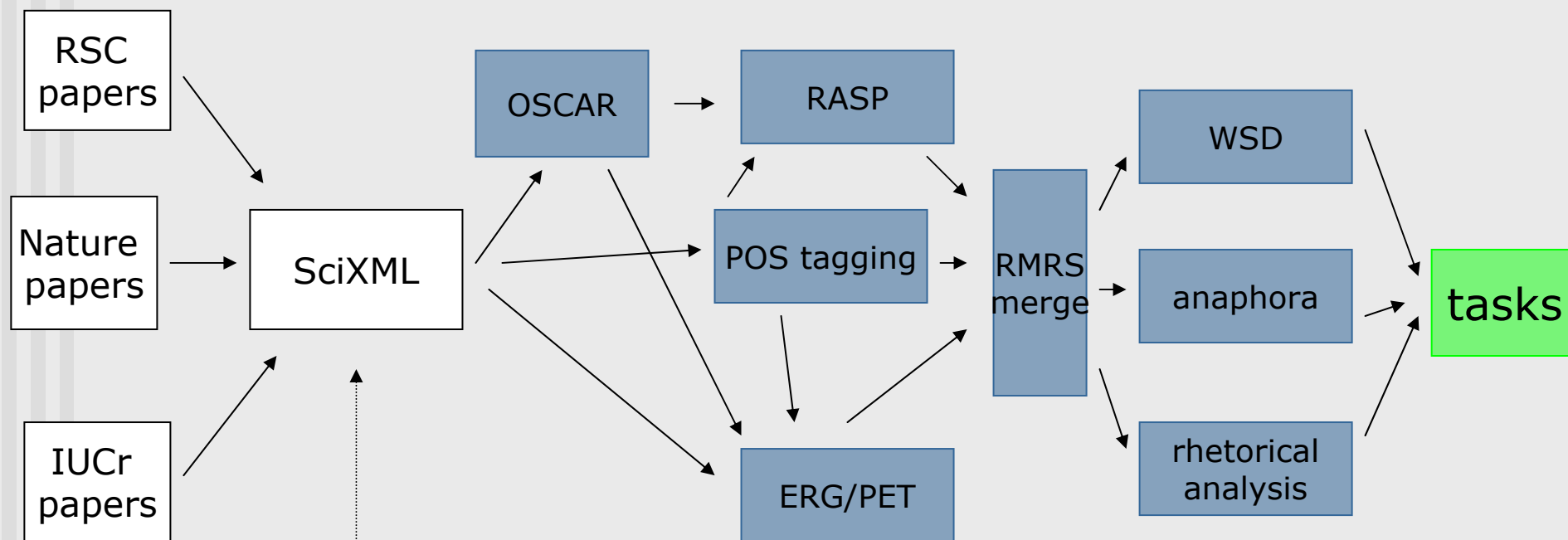
X cites Y (contrast)

# General Approach

---

- Integrate, adapt and further develop general tools for language processing (avoid domain-specific solutions wherever possible, even if immediate performance suffers)
- Incorporate deeper syntactic and compositional semantic processing (DELPH-IN technology)
  - Deep and shallow processing combined via semantics (RMRS)
- Integration with XML
  - SciXML standard for papers
  - All processing delivers standoff annotation (SAF/SMAF)
  - Architecture allows ambiguity to be preserved
- Links to semantic web ontologies to provide domain semantics
- Open Source where possible, collaborative development
- Eventual multilingual work via collaboration

# Outline architecture



standoff annotation

# SciXML

```
<?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>
```

# SciXML

<BODY>

<DIV DEPTH="1"><HEADER>Introduction</HEADER>

<P>Although the first Tröger's base <XREF ID="chem1" TYPE="COMPOUND">1</XREF> was obtained more than a century ago from the reaction of <IT>p</IT>-toluidine and formaldehyde,<REF ID="cit1" TYPE="P">1</REF> 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,<REF ID="cit2" TYPE="P">2</REF> chelating and biomimetic systems,<REF ID="cit3" TYPE="P">3</REF> and transition metal complexes for regio- and stereoselective catalytic reactions.<REF ID="cit4" TYPE="P">4</REF> For these reasons, numerous Tröger's-base derivatives have been prepared bearing different types of substituents and structures (<IT>i.e.</IT>, <XREF ID="chem2 chem3 chem4 chem5" TYPE="COMPOUND">2-5</XREF> Scheme 1), with the purpose of increasing their potential applications.<REF ID="cit2 cit3 cit5" TYPE="P">2,3,5</REF> 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 <XREF ID="chem4" TYPE="COMPOUND">4</XREF> and <XREF ID="chem5" TYPE="COMPOUND">5</XREF>.<REF ID="cit5e" TYPE="P">5<IT>e</IT></REF></P>

# SciXML

```
<REFERENCELIST>
```

```
<REFERENCE ID="cit1"><AUTHORLIST><AUTHOR>J.<SURNAME>Tröger</SURNAME></AUTHOR>
  </AUTHORLIST>
  <TITLE/>
  <JOURNAL><NAME>J.Prakt.Chem.</NAME>
    <YEAR>1887</YEAR><VOLUME>63</VOLUME>
    <PAGES>225</PAGES></JOURNAL>
</REFERENCE>
```

```
<REFERENCE
  ID="cit2"><AUTHORLIST><AUTHOR>M.D.<SURNAME>Coward</SURNAME></AUTHOR>
  <AUTHOR>I.<SURNAME>Sucholeiki</SURNAME></AUTHOR>
  <AUTHOR>R. R.<SURNAME>Bukownik</SURNAME></AUTHOR>
  <AUTHOR>C. S.<SURNAME>Wilcox</SURNAME></AUTHOR>
  </AUTHORLIST>
  <TITLE/>
  <JOURNAL><NAME>J. Am. Chem. Soc.</NAME>
    <YEAR>1988</YEAR><VOLUME>110</VOLUME>
    <PAGES>6204</PAGES></JOURNAL>
</REFERENCE>
```

# Standoff annotation (SMAF)

---

- Document markup may not be relevant for sentence parsing but must be preserved for later processing
  - e.g., section and paragraph boundaries needed for anaphora resolution
- Preserve ambiguity of processing stages as necessary (via lattice)
- Grounding with respect to original text/SciXML (via character position or XPath plus Xpointer)
- Generic framework for standoff pointers and ambiguity, different content specification for different levels (sentences, tokens, feature structures, RMRS)
- SMAF proposal (Waldron et al, LREC 2006), derived from MAF (ISO draft) and Pet XML Input Chart (PIC).
- Refined version of Heart of Gold approach developed for Deep Thought project (Schäfer et al)

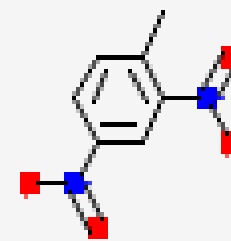
# Chemistry names: OSCAR

## 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**

# Parsing

---

- RASP
  - 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, reasonably fast, relatively shallow, no conventional semantics in output
- LKB/PET plus ERG (English Resource Grammar)
  - DELPH-IN [www.delph-in.net](http://www.delph-in.net)
  - Flickinger, Oepen, Copestake, Callmeier et al
  - HPSG, stochastic ranking
  - detailed lexicon, POS tagging for unknown words
  - missing lexicon causes problems, relatively slow (though exceedingly fast by normal HPSG standards!), detailed semantic output in Minimal Recursion Semantics (MRS)

# Integrating processing in general

---

- No single system can do everything: deep and shallow processing have inherent strengths and weaknesses
  - shallow: speed and robustness: e.g., POS tagging, chunking
  - deep: detail, precision, potential for bidirectional processing: e.g., HPSG-based parsers and generators (DELPH-IN technology)
  - also intermediate: RASP (Robust accurate statistical parser): relatively detailed but no lexicon.
- Domain-dependent and domain-independent processing must be linked
- Desirable to have a common representation language for processing **above sentence level** (e.g., anaphora)
- We use an underspecified semantics for that purpose (RMRS)

# RMRS: Extreme underspecification

---

- Minimal Recursion Semantics: MRS. Compositional semantics for deep processing (Copestake, Flickinger, Sag and Pollard, 1999, in press)
  - scope underspecification
  - hierarchies for predicates
  - accumulate information monotonically by simple operations
  - adopted for DELPH-IN and other HPSG work, also compatible with LFG etc
- Robust MRS (RMRS): adaptation of MRS allowing shallower processing
  - Split up semantic representation into minimal components
    - split up predicate argument structure
  - Don't represent what you don't know but preserve everything you do know at each processing stage
  - Flat representation allows pieces of the semantics to be accessed individually (**semantics on packed representations**)
  - MRS can be converted to RMRS

# POS output as underspecification

“Every cat chased some dog”

DEEP –

```
lb1: _every_q(x1sg), RSTR(lb1,h9), BODY(lb1,h6),  
lb2: _cat_n(x2sg), lb5: _dog_n_1(x4sg),  
lb4: _some_q(x3sg), RSTR(lb4,h8),  
BODY(lb4,h7), lb3: _chase_v(esp),  
ARG1(lb3,x2sg), ARG2(lb3,x4sg), h9=lb2,h8=lb5,  
x1sg=x2sg, x3sg=x4sg
```

POS –

```
lb1: _every_q(x1), lb2: _cat_n(x2sg),  
lb3: _chase_v(epast), lb4: _some_q(x3),  
lb5: _dog_n(x4sg)
```

# POS output as underspecification

“Every cat chased some dog”

DEEP –

```
lb1: _every_q(x1sg), RSTR(lb1,h9), BODY(lb1,h6),
lb2: _cat_n(x2sg), lb5: _dog_n_1(x4sg),
lb4: _some_q(x3sg), RSTR(lb4,h8),
BODY(lb4,h7), lb3: _chase_v(esp),
ARG1(lb3,x2sg), ARG2(lb3,x3sg), h9=lb2,h8=lb5,
x1sg=x2sg, x3sg=x4sg
```

POS –

```
lb1: _every_q(x1), lb2: _cat_n(x2sg),
lb3: _chase_v(epast), lb4: _some_q(x3),
lb5: _dog_n(x4sg)
```

# RMRS construction

---

- deep grammars: MRS  $\leftrightarrow$  RMRS converter
- OSCAR: 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
  - output aims to match deep grammar (ERG)
  - developed on basis of ERG semantic test suite
  - default composition principles when no rule RMRS specified
- Research Markup: RMRS versions of cue phrases

# 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

Also, we have previously reported the synthesis of pyrazole and pyridine Tröger's base analogues. In this work, we report the synthesis of pyrazole and pyridine Tröger's base analogues. In this work, we report the synthesis of pyrazole and pyridine Tröger's base analogues.

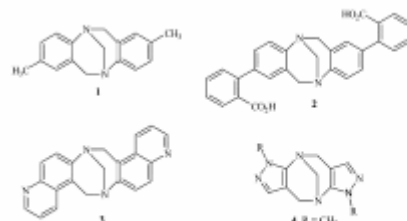
Rodrigo Abonia, Andrea Alborno, Hector Larrahondo, Jairo Quiroga, Braulio Insuasty, Henry Insuasty, Angelina Hormaza, Adolfo Sánchez, Manuel Nogueras

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-arylpurazoles 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 <sup>1</sup>H and <sup>13</sup>C NMR, mass spectra and elemental analysis and confirmed by X-ray diffraction for one of the obtained compounds.

metal clusters of some of the group eight metals (*i.e.*, Fe, Ru and Os), as re

## Introduction

Although the first Tröger's base **1** was obtained more than a century ago from the reaction of *p*-toluidine and formaldehyde,<sup>[1]</sup> 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,<sup>[2]</sup> chelating and biomimetic systems,<sup>[3]</sup> and transition metal complexes for regio- and stereoselective catalytic reactions.<sup>[4]</sup> 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.

increasing their potential applications.<sup>[2,3,5]</sup> 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-arylpurazoles **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 quinoline **7a**, which could be used as an intermediate in the synthesis of new hydroquinoline analogues of interest,<sup>[6]</sup> 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-

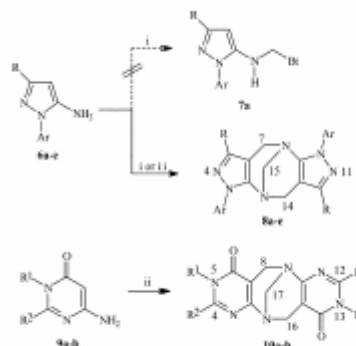


Chart 1 Reaction of 5-amino-3-methyl-1-phenylpyrazole **6a** and 6-aminopyrimidin-4(3*H*)-one **9** with formaldehyde. Reagents: i = CH<sub>2</sub>O, BH, EtOH, HOAc; ii = CH<sub>2</sub>O, EtOH, HOAc, BH = Benzotriazole.

compounds **8a–e** and **10a,b** by u

pentadeca-2(6),4,9(13),11-tetraene <XR

Also, we have previously reported the synthesis of pyrazole and pyrimidine Tröger's base analogues in mixtures (i.e., compounds 11e<

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

Tröger's-base analogues bearing fused pyrazolic or pyrimidin 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).

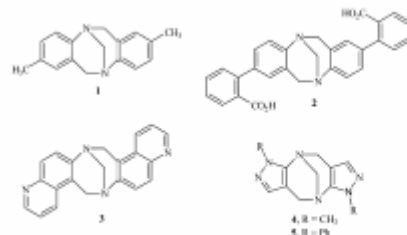
Two key intermediates were isolated from the reaction mixtures, which helped us to suggest a sequence of steps for the formation of the Tröger's bases obtained. The structures of the products were assigned by <sup>1</sup>H and <sup>13</sup>C NMR, mass spectra and elemental analysis and confirmed by X-ray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

metal clusters of some of the group eight metals (i.e., Fe, Ru and Os), as macrocyclic organic molecules. (i.e., cit12" TYPE="P">12</REF></P></DIV>

## 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-

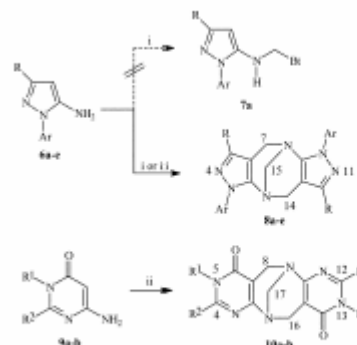


Chart 1 Reaction of 5-amino-3-methyl-1-phenylpyrazoles **6** and 6-aminopyrimidin-4(3*H*)-ones **9** with formaldehyde. Reagents: i = CH<sub>2</sub>O, BH = HOAc, HOAc; ii = CH<sub>2</sub>O, EtOH, HOAc; BH = Benzotriazole.

DOI: 10.1039/b200862a

## Legenda:

Background

Other

Own

Based

Contrast

Textual

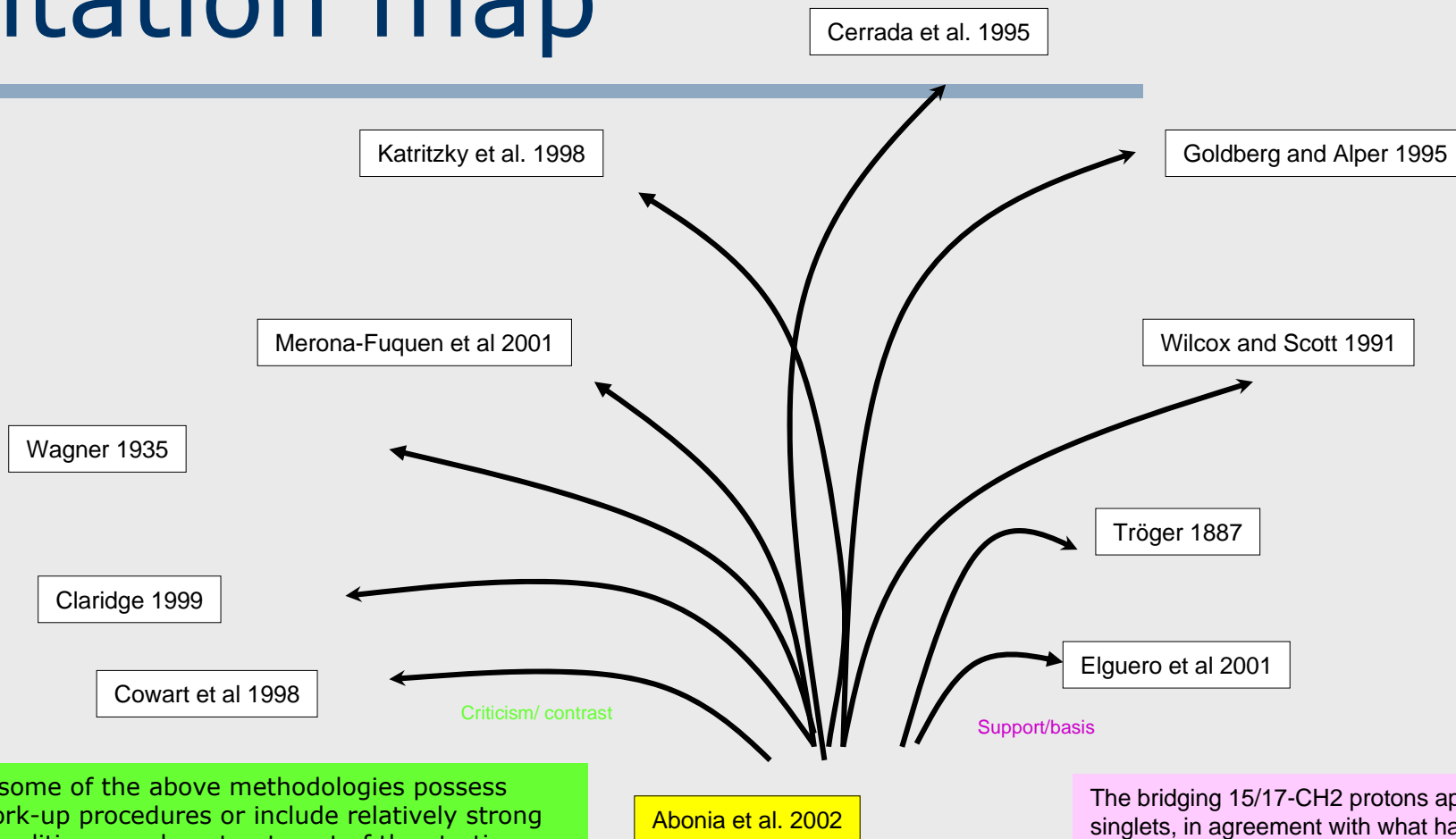
Aim

# Argumentative Zoning

- Robust discourse structure for scientific articles
- 7 flat labels
- Automatically recognisable from annotated material; relatively shallow automatic features, recognition in range of  $K=.5$  (humans:  $K=.7$ )
- Principles:
  - Attribution of authorship: who does what?
    - Passives resolved by context and position in paper
  - Author stance towards cited work
  - General goal statements:

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.

# 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-CH<sub>2</sub> protons appear as singlets, in agreement with what has been observed for similar systems [9].

# Citation relation recognition

---

- Types of information recognisable:
  - Contradictory claims, as opposed to confirmation of previous claims
  - “paradigm shift” sentences: Contrary to received wisdom...
  - Advantages over rival methods
  - Usage of earlier, cited methods and procedures
- Annotation scheme for citation function
  - 14 categories, eg. “Uses Definition introduced in cited source”
  - Humans are currently being trained to apply it
  - More subjectivity involved than in Argumentative Zoning
- Automatic recognition: Use similar surface features

# Citation Function: Annotation scheme

---

**Weak** (3%) – cited work has a weakness

**PUse** (15%)– uses data or algorithm from cited work (unchanged)

**PModi** (2%)– uses data or algorithm from cited work (changed)

**PBas** (2%) – bases itself on cited work

**PDef** (1.5%)– uses definition in cited work

**PSup** (2%) - support for or from cited work

**PSim** (4%) – similarity to cited work

**PMot** (2%) – motivates: approach works, or problem addressed relevant

**CoCoM** (8%) – contrast in method

**CoCoG** (0.5%)– contrast in goal

**CoCoRO** (5%) – contrast in results (neutral or worse than cited work)

**CoCoR-** (2%) - contrast in results, cited work is worse

**CoCoXY** (2%)– contrast between two OTHER cited works

**Neut** (51%)– Neutral (or not enough textual evidence)

# Features for AZ and citation analysis

---

- Type of verb
  - 'problem' verb: fail, not\_manage, have\_problem\_with, run\_into\_problem
  - 'presentation' verb: show, present, propose
- Type of agent/subject
  - We/our system/this paper
  - Them/their approach/this result
- Indicator phrases
  - However, their approach cannot... (criticism)
  - Whereas they ..., we... (contrast)
- Citation type and location
  - Self-citation or not?
  - Beginning or end of sentence?
- Low-level features:
  - Relative and absolute location in article/paragraph
  - Sentence density of TF\*IDF words
  - Overlap of sentence with headline
  - Headline type, sentence length...

# RMRS and research markup

---

- Specify cues in RMRS: e.g.,
  - **l1:objective(x), ARG1(l1,y), l2: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

---

- Project SciBorg:
  - NL markup language with link to Web languages platform for extraction of information, glue language (RMRS)
  - IE technology and ontologies (ie Oscar for parsing of chemistry compound names)
  - Modelling scientific argumentation and citation purpose for novel ways of info access
  - Real-world e-science environment