



UNIVERSITY OF
CAMBRIDGE

Integrating General-Purpose and Domain-Specific Components in the Analysis of Scientific Text

CJ Rupp

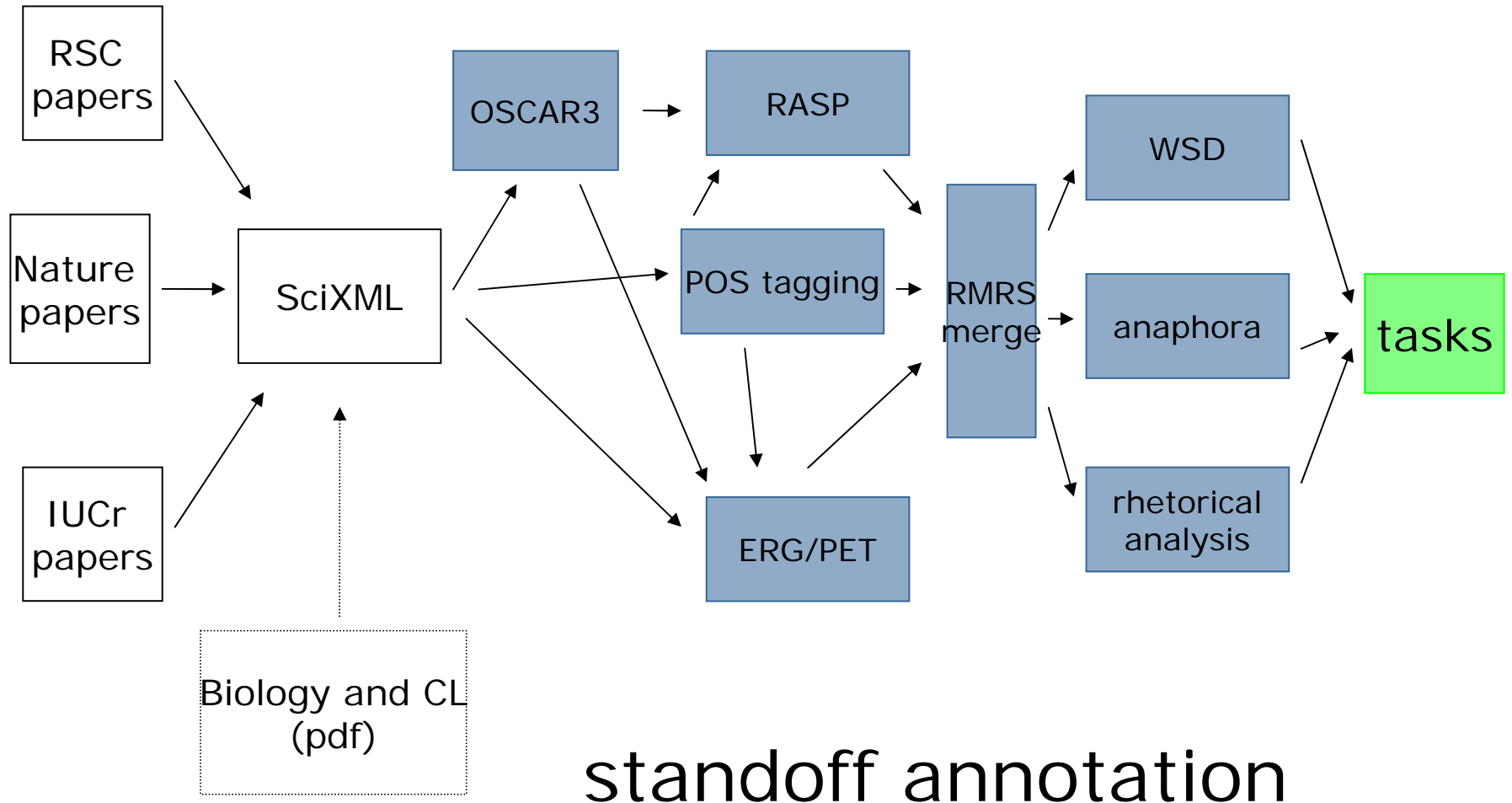
*Ann Copestake, Peter Corbett,
Benjamin Waldron*

University of Cambridge Computer Laboratory
Email: cr351@cl.cam.ac.uk

The SciBorg Project

- An eScience project for Information Extraction from published chemistry research papers.
- The texts are supplied by journal publishers in XML markup.
- We use multiple analysis engines for maximum coverage.
- Analyses stored as standoff (SAF) annotations, either in a database or XML.
- RMRS (Robust Minimal Recursion Semantics) for common representations.

SciBorg Architecture



Chemistry Research Text...

- ... looks like this:

Dialkyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylates are obtained in excellent yields from the 1 : 1 : 1 addition reaction between triphenylphosphine, dialkyl acetylenedicarboxylates and 3-chloroindole-2-carbaldehyde; dimethyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate is converted to dimethyl 9-oxo-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate.

- Whereas *real* text looks more like:

This window contains an HP-UX shell (either a Bourne shell or C-shell depending on the value of the SHELL environment variable; for details see the “Concepts” section of the “Using Commands” chapter).

Efficient synthesis of functionalized 3*H*-pyrrolo[1,2-*a*]indoles

Issa Yavari,* Mehdi Adib and Mohammad H. Sayahi

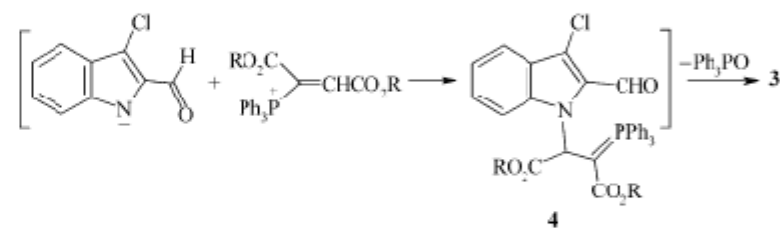
Department of Chemistry, University of Tarbiat Modarres, P. O. Box 14115-175, Tehran, Iran.
E-mail: isayavar@yahoo.com; Fax: +98 21 8006544

Received (in Cambridge, UK) 8th May 2002, Accepted 28th May 2002
First published as an Advance Article on the web 10th June 2002

Dialkyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylates are obtained in excellent yields from the 1 : 1 : 1 addition reaction between triphenylphosphine, dialkyl acetylenedicarboxylates and 3-chloroindole-2-carbaldehyde; dimethyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate is converted to dimethyl 9-oxo-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate.

Introduction

Bridgehead nitrogen heterocycles are of interest because they constitute an important class of natural and non-natural products, many of which exhibit useful biological activity.¹⁻⁴ The interest in tricyclic-fused 5:5:6 systems with one ring junction nitrogen atom between the five-membered rings and no extra heteroatoms, stems from the appearance of saturated and partially saturated annelated[*a*]indole ring systems in many biologically active compounds. Consequently, there has been an ongoing interest in the synthesis of pyrrolo[1,2-*a*]indole ring structures, especially synthesis of 2,3,9,9a-tetrahydro-5,8-dioxo-1*H*-pyrrolo[1,2-*a*]indole, the “mitosane”, basic skeleton of mitomycins.³⁻⁷ With the purpose to prepare pyrrolo[1,2-*a*]indole derivatives, we now report the reaction of 3-chloroindole-2-carbaldehyde **1** and dialkyl acetylenedicarboxylates **2**



Scheme 2

compounds displayed molecular ion peaks at $m/z = 305$, 333 and 389 , respectively. The ^1H NMR spectrum of compound **3a** exhibits two single sharp lines for the methoxy ($\delta = 3.78$ and 3.87 ppm) protons. The two non-aromatic methine protons appear as two doublets at $\delta = 5.57$ and 7.65 ppm with allylic coupling of $^4J_{\text{HH}} = 1.9$ Hz. The ^{13}C NMR spectrum of **3a** exhibits a signal at $\delta = 63.75$ ppm for the N-CH moiety. The ^1H and ^{13}C NMR spectra of **3b** and **3c** are similar to those of **3a**, except for the ester moieties, which exhibited characteristic resonances with appropriate chemical shifts.

When compound **3a** was refluxed in boiling toluene for 24 h, the ^1H NMR spectrum of the reaction mixture showed quantitative conversion to dimethyl 9-chloro-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate (**5**). Compound **5** was quantitatively hydrolyzed to dimethyl 9-hydroxy-9*H*-pyrrolo[1,2-*a*]indole-2,3-

General procedure for preparation of compounds 3a–c

Dimethyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate (3a). The typical process for the preparation of dimethyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate (3a) is described as an example. To a magnetically stirred solution of 0.524 g triphenylphosphine (2 mmol) and 0.359 g 3-chloroindole-2-carbaldehyde **3** (2 mmol) in 6 mL dichloromethane was added dropwise a mixture of 0.284 g dimethyl acetylenedicarboxylate (2 mmol) in 2 mL dichloromethane at $-5\text{ }^{\circ}\text{C}$ over 10 min. The reaction mixture was then allowed to warm up to room temperature and stirred for 15 min. The solvent was removed under reduced pressure and the product was extracted from the solid residue with $4 \times 20\text{ mL}$ *n*-hexane. The solvent was removed under reduced pressure and the product was recrystallized from *n*-hexane–ethyl acetate (1 : 1) as yellow crystals, mp $136\text{--}139\text{ }^{\circ}\text{C}$, yield 0.6 g, 98%. IR (KBr) ($\nu_{\text{max}}/\text{cm}^{-1}$): 1731 and 1693 (C=O). MS, m/z (%): 305 (M^+ , 45), 290 (10), 270 (5), 246 (100), 216 (8), 187 (20), 152 (12). Anal. Calcd for $\text{C}_{15}\text{H}_{12}\text{NO}_4\text{Cl}$ (305.72): C, 58.93; H, 3.96; N, 4.58. Found: C, 58.9; H, 4.0; N, 4.6%. ^1H NMR: δ 3.78 and 3.87 (6 H, 2 s, 2 OCH_3), 5.57 (1 H, d, $^4J = 1.9\text{ Hz}$, CH), 7.18 (1 H, t, $J = 8.0\text{ Hz}$, CH), 7.29 (1 H, t, $J = 7.9\text{ Hz}$, CH), 7.35 (1 H, d, $J = 8.2\text{ Hz}$, CH), 7.64 (1 H, t, $J = 8.1\text{ Hz}$, CH), 7.65 (1 H, d, $^4J = 1.9\text{ Hz}$, CH). ^{13}C NMR: δ 52.23 and 53.28 (2 OCH_3), 63.75 (CH), 101.60 (C), 110.03, 120.10, 120.97, and 125.07 (4 CH), 129.79 (C), 130.89 (CH), 133.94, 134.57 and 139.34 (3 C), 162.58 and 167.11 (2 C=O).

2 OCH_3), 5.85 (1 H, s, CH), 6.77 (1 H, s, CH), 7.27 (1 H, t, $J = 7.6\text{ Hz}$, CH), 7.40 (1 H, t, $J = 7.8\text{ Hz}$, CH), 7.56 (1 H, d, $J = 7.5\text{ Hz}$, CH), 7.83 (1 H, d, $J = 8.1\text{ Hz}$, CH). ^{13}C NMR: δ 49.17 (CH), 51.93 and 52.72 (2 OCH_3), 108.13 and 114.63 (2 CH), 120.97 and 124.21 (2 C), 126.26, 126.62, and 130.45 (3 CH), 136.34, 138.36, and 139.18 (3 C), 161.58 and 164.06 (2 C=O, ester).

Dimethyl 9-hydroxy-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate (6)

0.30 g (1 mmol) of **5** was refluxed in a mixture of chloroform (10 mL) and water (2 mL) for 12 h. Then 20 mL chloroform was added. The organic layer was dried over anhydrous sodium sulfate. The solvent was removed under reduced pressure and the solid residue was recrystallized from *n*-hexane–ethyl acetate (1 : 2) as colorless crystals, mp $159\text{--}161\text{ }^{\circ}\text{C}$, yield 0.28 g, 99%. IR (KBr) ($\nu_{\text{max}}/\text{cm}^{-1}$): 3425 (OH), 1709 and 1685 (C=O). MS, m/z (%): 287 (M^+ , 82), 270 (100), 228 (15), 216 (17), 59 (6). Anal. Calcd for $\text{C}_{15}\text{H}_{13}\text{NO}_5$ (287.27): C, 62.72; H, 4.56; N, 4.88. Found: C, 62.8; H, 4.5; N, 4.9%. ^1H NMR: δ 3.06 (1 H, d, $J = 9.7\text{ Hz}$, OH), 3.72 and 3.93 (6 H, 2 s, 2 OCH_3), 5.58 (1 H, d, $J = 9.7\text{ Hz}$, OH) 6.65 (1 H, s, CH), 7.22 (1 H, t, $J = 7.5\text{ Hz}$, CH), 7.34 (1 H, t, $J = 7.8\text{ Hz}$, CH), 7.55 (1 H, d, $J = 7.5\text{ Hz}$, CH), 7.74 (1 H, d, $J = 8.1\text{ Hz}$, CH). ^{13}C NMR: δ 51.84 and 52.56 (2 OCH_3), 67.12 (CHOH), 107.42 and 114.36 (2 CH), 120.34 and 123.59 (2 C), 125.97, 126.02 and 129.87 (3 CH), 138.33, 139.29 and 140.97 (3 C), 161.72 and 164.51 (2 C=O, ester).

Dimethyl 9-oxo-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate (7)

Chemistry and English

- We prefer to treat the text of research papers as English with some perturbation, *i.e.* chemistry,
- But:
 - *The density varies, some sections are immediately readable, others are virtually tables as text.*
 - *The structures are essentially English with unfamiliar terms and notations.*
 - *So the perturbation goes down to the tokenizer level.*
- *We need a separate analysis tool for the chemistry*

OSCAR3

- A system designed to recognise chemical terms in text and assign chemical structures.
- We don't need the structure assignments but 5 essential functions are distinguished:
 - Compounds : citric acid, dialkylpyridines, $C_6H_{12}O_6$
 - Chemical adjectives: citric, pyrazolic, aqueous
 - Reactions: dihydroxilate, iodise, chlorinated
 - Enzymes: methylase, nitrogenase, ethyltransferase
 - Chemical Prefixes: 1,3-dipolar, cis-isomer, α -position

OSCAR3 Categories

- The vast majority of chemical terms are nominals, including the compound category (CM) and enzymes (ASE).
- There is a small number of chemical adjectives (CJ).
- Reactions (RN) can correspond to virtually any open class category.
- Chemical prefixes (CPR) may also appear as isolated tokens.

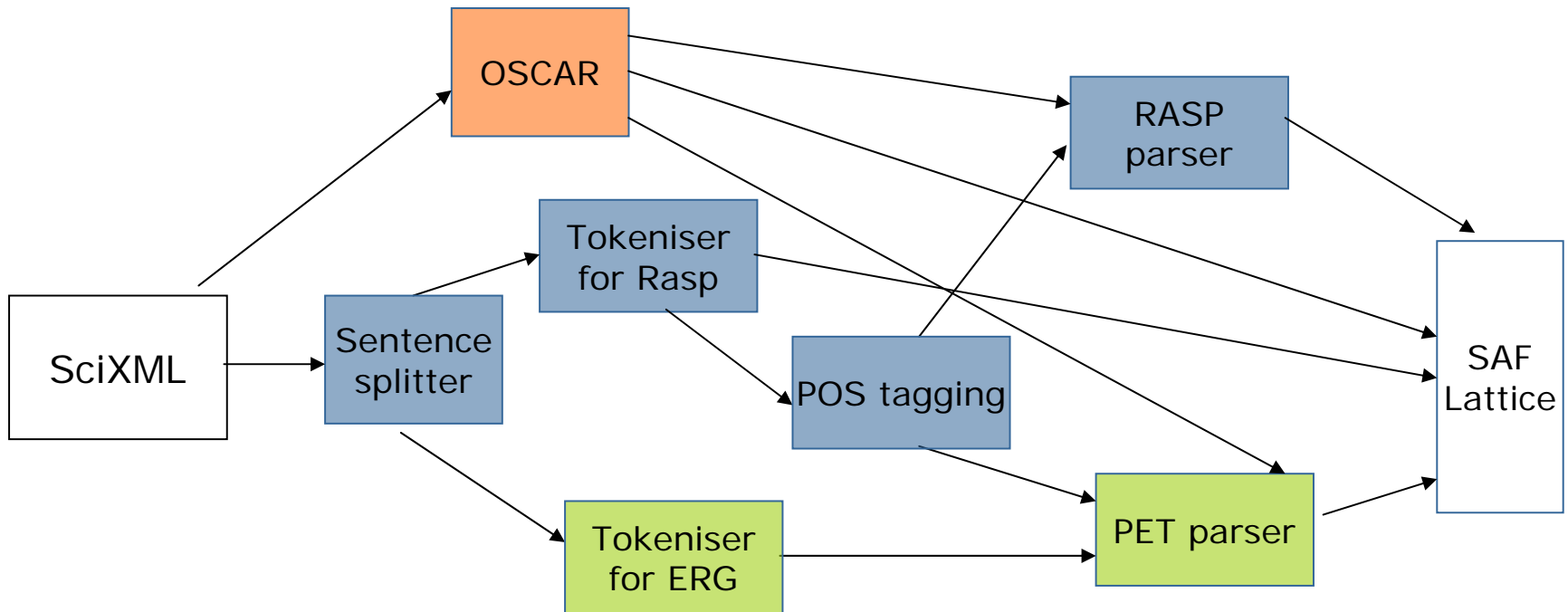
SciBorg Framework

- Three important representations:
 - SciXML a common XML markup sharing the necessary features of the publishers' in-house markup formats.
 - SAF standoff annotations for each level of analysis: sentence, token, tag, RMRS and OSCAR3 annotations.
 - RMRS underspecified semantics.
- Analysis maps from SciXML to RMRS.

Multi-Engine Analysis

- PET/ERG: “deep” analysis using detailed (HPSG) grammars and lexicons.
- RASP: Robust shallow parsing with a statistically trained grammar.
- Key phases in each analysis strand are factored into separate modules: e.g. tokeniser, parser,...
- OSCAR3 analyser included as a single module.
- Results on different strands can recombine to enhance robustness, i.e. not a strict pipeline.

SciBorg Parsing Architecture



RASP Parser

- In the SciBorg architecture the RASP parser comprises 4 modules.
 - The sentence splitter is a flex script.
 - The tokeniser is also in flex.
 - The tagger can return multiple weighted tags.
 - The parser is a statistical CFG parser that can return RMRS, by a tree walking conversion.
- The parser and tagger are trained on the SUSANNE corpus of English.

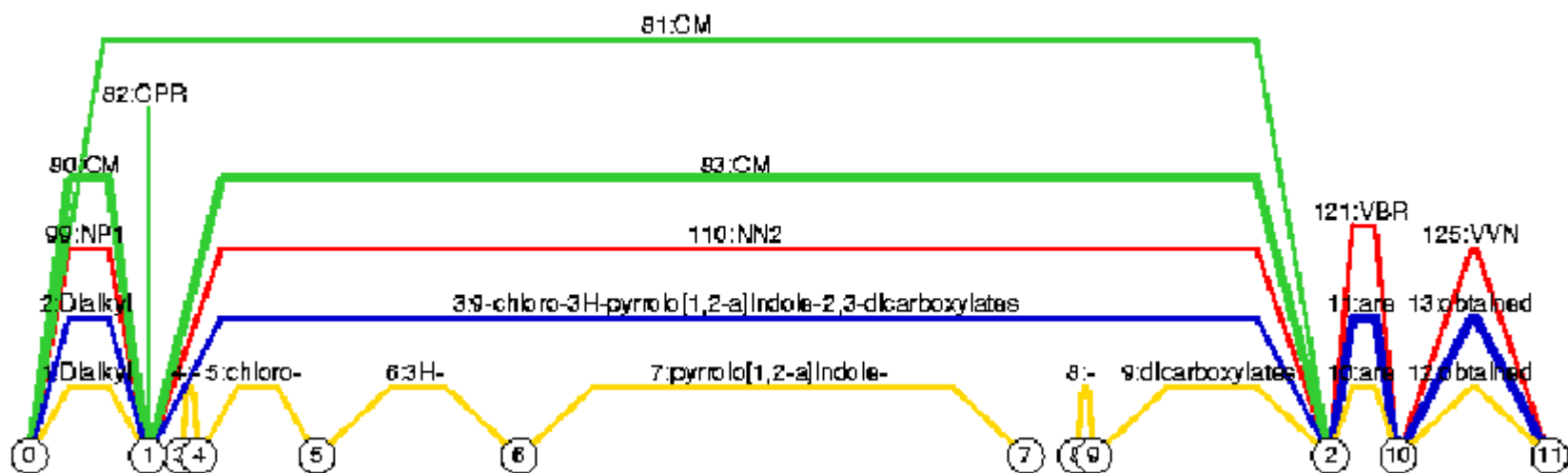
ERG Analysers

- The PET parser and FSPP tokeniser factored into separate modules.
- There is also an LKB module for testing.
- RASP tagging results are used as a basis for the PET unknown word mechanism.
- Parser inputs are encoded in a SMAF lattice to incorporate OSCAR3 and FSPP tokenisations.

Integration

- We get multiple results at different levels.
- The SAF lattice ensures that they are all available.
- There's always something to fall back on, but are you always getting the right or, at least, the best results.
- In general, you want to favour the OSCAR3 analyses.

A Lattice of SAF Annotations



FSPF-tokenizer, OSCAR3

RASP tagger, RASP tokenizer

Approximation in ERG Parsing

- OSCAR3 categories map to general lexical types:

```
oscar.[type='CM'] -> gMap.type='n_-_pn-unk_le'  
                    tokenStr='OSCARCOMPOUND'
```

```
oscar.[type='CM'] -> gMap.pred='chem_compound_rel'  
                    gMap.carg=content.SMILES
```

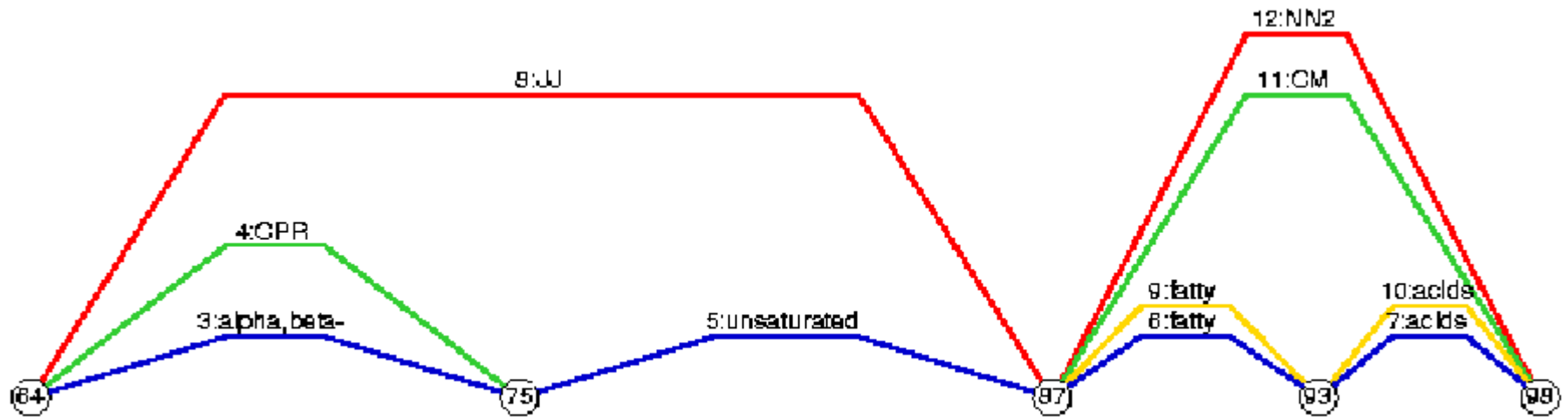
- As the functional information is limited
- RASP tags are also mapped to lexical types:

```
pos.[tag='NN1'] -> gMap.type='n_-_c-sg-unk_le'  
pos.[tag='JJ'] -> gMap.type='aj_-_i-unk_le'
```

Specific Problems

- OSCAR3 term that doesn't correspond to any parser token, e.g. prefix (CPR).
- OSCAR3 function not integrated.
 - If we take RASP we may get a span that ignores the chemistry.
 - If we take ERG we cannot ensure an analysis with OSCAR3 tokens is preferred.
- Just inconsistent segmentation.

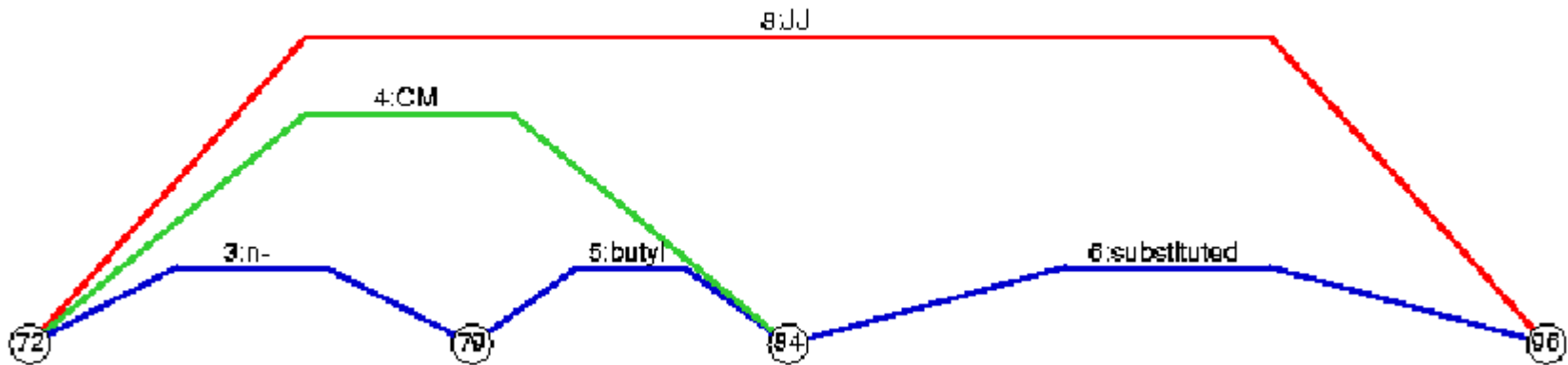
α,β -unsaturated fatty acids



FSPF-tokenizer, OSCAR3

RASP tagger, RASP tokenizer

n-butyl-substituted



FSPF-tokenizer, OSCAR3

RASP tagger

What Next?

- We're retraining the parse selection models to include chemical terms.
 - Can we use external confidence scores?
- We could handle chemistry prefixes (CPR) as syntactically null items.
 - Might not work in coordinate structures.
- Other types of terms are recognisable:
 - Citations, measure phrases.

SciXML

- This is a markup schema for research papers.
- It preserves the necessary features of the publishers' in-house markup.
- The main organising principle is the logical structure of the paper.
- SciBorg also requires the rendering of examples to be recognisable to the user.

Underspecified Semantics

- Put simply: *if you don't have enough information to resolve things then wait.*
- Typically this involves representing information about a set of logical formulae.
- RMRS (Robust Minimal Recursion Semantics) is the most active underspecification framework.
- Robustness comes from leavin

Text out of XML Markup

- Only some sections of a marked up research paper contain text:
 - E.g. <P> , <ABSTRACT> , <CAPTION>
- With common markup we can just list the elements with text in them.
- But what about the inline markup?
 - *The mixture was stirred for 2 h at rt whereupon H<SB>2</SB>O (10 cm<SP>3</SP>) was added.*

XML Markup out of Text

- There's basically three things you can do with an element:
 - Process it: see text elements above
 - Skip it: cross references, footnotes
 - Ignore it: inline font markup `<SB>` , `<IT>`
- So we have a patchwork of fragments of the XML file that constitute the text.

General Issues

- The logical structure of research text
 - Citations
 - Cross references
- Including markup in analysis?
 - Emphasis?
- Utilising external weightings
 - Tagger scores
 - OSCAR3 confidences

Character Offset Indexing

Formatted text: Come *here*!

raw text: "<p>Come <i>here</i>!</p>"

Unicode character points:

.<.p.>.C.o.m.e. .<.i.>.h.e.r.e .< ./ .i .> .! .< ./ .p .>

·
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
23

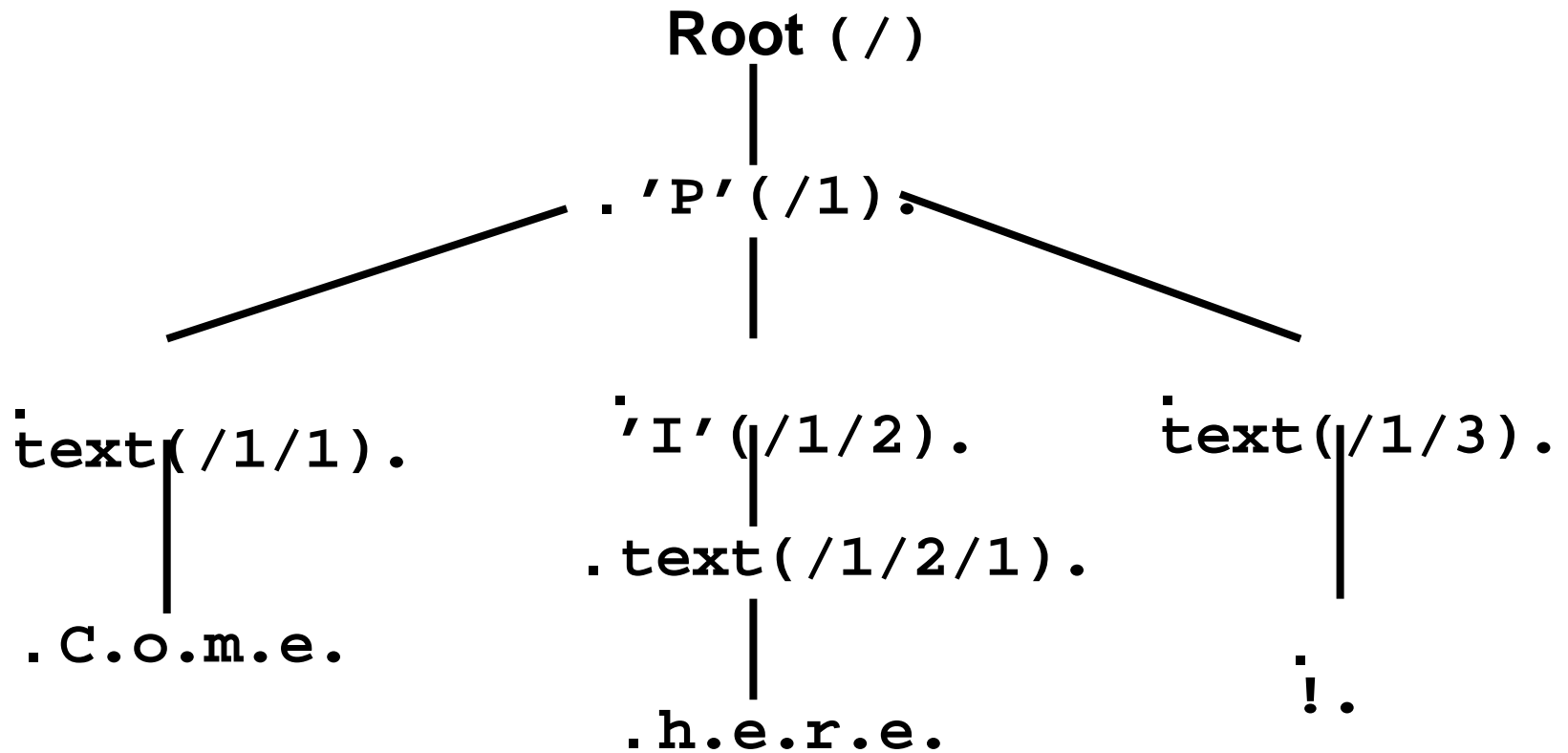
Tokens

<token from='3' to='7' value='Come' />

<token from='11' to='14' value='here' />

<token from='18' to='19' value='!' />

XPoint Indexing



Types of SAF Annotation

- Sentence segments
 - `<annot type='sentence' id='s133' from='42065' source='v4987' target='v5154' to='43039' value='...calculated for C11H18O3....'/>`
- Tokens
 - `<annot type='token' id='t5151' from='42988' to='43030' deps='s133' source='v5150' target='v5151' value='calculated'/>`
 - `<annot type='token' id='t5152' from='43031' to='43034' deps='s133' source='v5151' target='v5152' value='for'/>`
 - `<annot type='token' id='t5153' from='43035' to='43043' deps='s133' source='v5152' target='v5153' value='C11H18O3'/>`

Types of SAF Annotation

- Part of Speech (POS) Tags
 - `<annot type='pos' id='p5151' deps='t5151' source='v5150' target='v5151' value='VFN'/>`
 - `<annot type='pos' id='p5152' deps='t5152' source='v5151' target='v5152' value='IF'/>`
 - `<annot type='pos' id='p5153' deps='t5153' source='v5152' target='v5153' value='NP1'/>`
- OSCAR (NER) mark up
 - `<annot from="/1/5/6/27/51/2/83.1" to="/1/5/6/27/51/2/88/1.1" type="oscar" id="o554"><slot name="type">compound</slot><slot name="surface">C11H18O3</slot><slot name="provenance">formulaRegex</slot></annot>`

Types of SAF Annotation

- RMRS analyses

```
<annot type='rmrs' id='r2' from='1029' source='v34' target='v45' to='1130'>
```

```
<rmrs cfrom='1029' cto='1130'>
```

```
<label vid='79'>
```

```
<ep cfrom='1029' cto='1129'><gpred>prpstn_m_rel</gpred><label vid='66'><var sort='h'  
vid='69'></ep>
```

```
<ep cfrom='1029' cto='1129'><gpred>bare_div_q_rel</gpred><label vid='70'><var  
sort='x' vid='2'></ep> . . .
```

```
. . . <hcons hreln='qeq'><hi><var sort='h' vid='14'></hi><lo><label  
vid='13'></lo></hcons>
```

```
<hcons hreln='qeq'><hi><var sort='h' vid='49'></hi><lo><label vid='23'></lo></hcons>
```

```
<hcons hreln='qeq'><hi><var sort='h' vid='34'></hi><lo><label vid='27'></lo></hcons>
```

```
</rmrs>
```

```
</annot>
```