Discourse Processing for Scientific Papers

Simone Teufel
Natural Language and Information Processing Group
Computer Laboratory

UNIVERSITY OF
CAMBRIDGE

NaCTeM seminar, Manchester, Jan 26 2007

“Scientific Processing” Projects at Cambridge

Citraz (EPSRC, PI Teufel, 2004-2007):

- Citation Maps and Citation Function Classification
- Domain: Computational Linguistics

FlySlip (BBSRC, PI Briscoe, 2005-2008):

- Part-automation of Curation of FlyBase (DB of Drosophila genes)
- GUI important aspect
- Use of statistical parser (RASP)

SciBorg (EPSRC, 2006-2009)

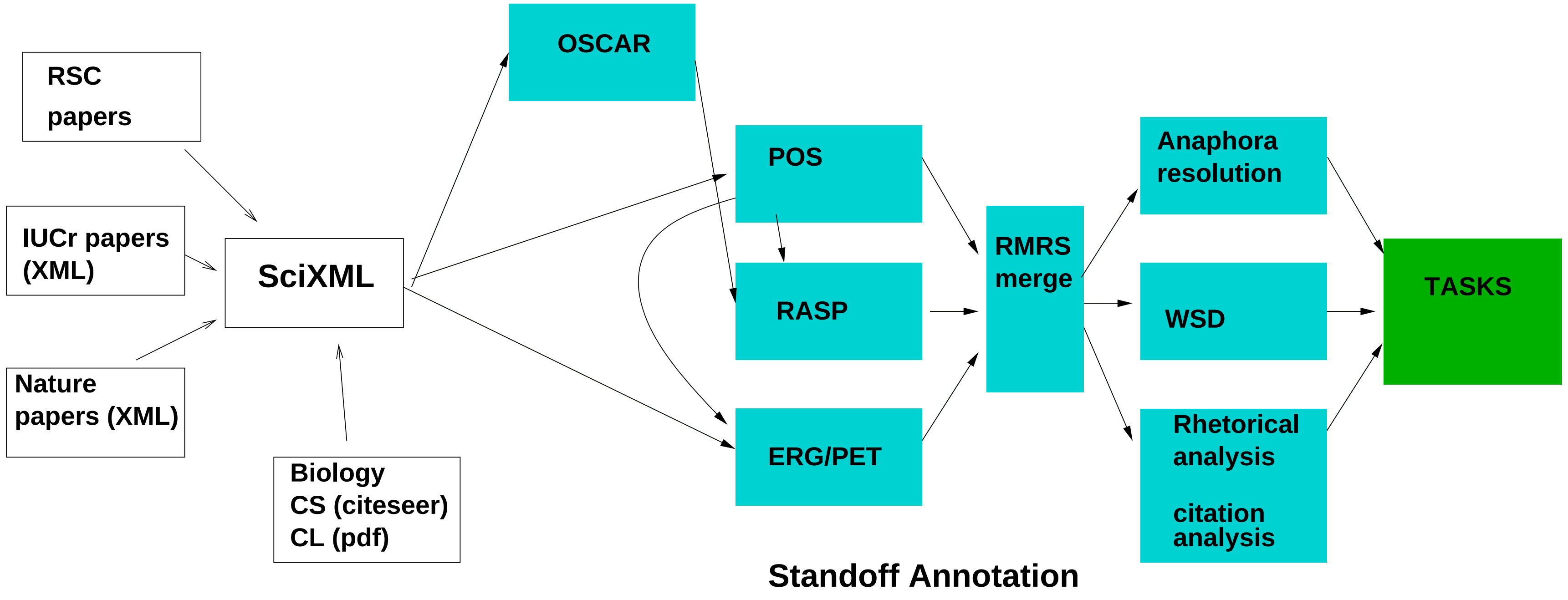
Project SciBorg: Knowledge Management for Chemists

- PIs Copestake, Teufel, Murray-Rust (Chemistry Dept Cambridge), Parker (CeSC)
- RAs CJ Rupp, Peter Corbett, Advaith Siddharthan
- Partners: Nature, Royal Society of Chemistry, International Union of Crystallography
- Aims:
  - Develop a NL markup language (RMRS) which acts as platform for IE. Link to semantic web languages.
  - Develop IE technology and ontologies for use by publishers, researchers, readers, vendors, and regulatory organisations
  - Model scientific argumentation and citation function for new ways of information access
  - Demonstrate applicability of this infrastructure in real-world eScience.

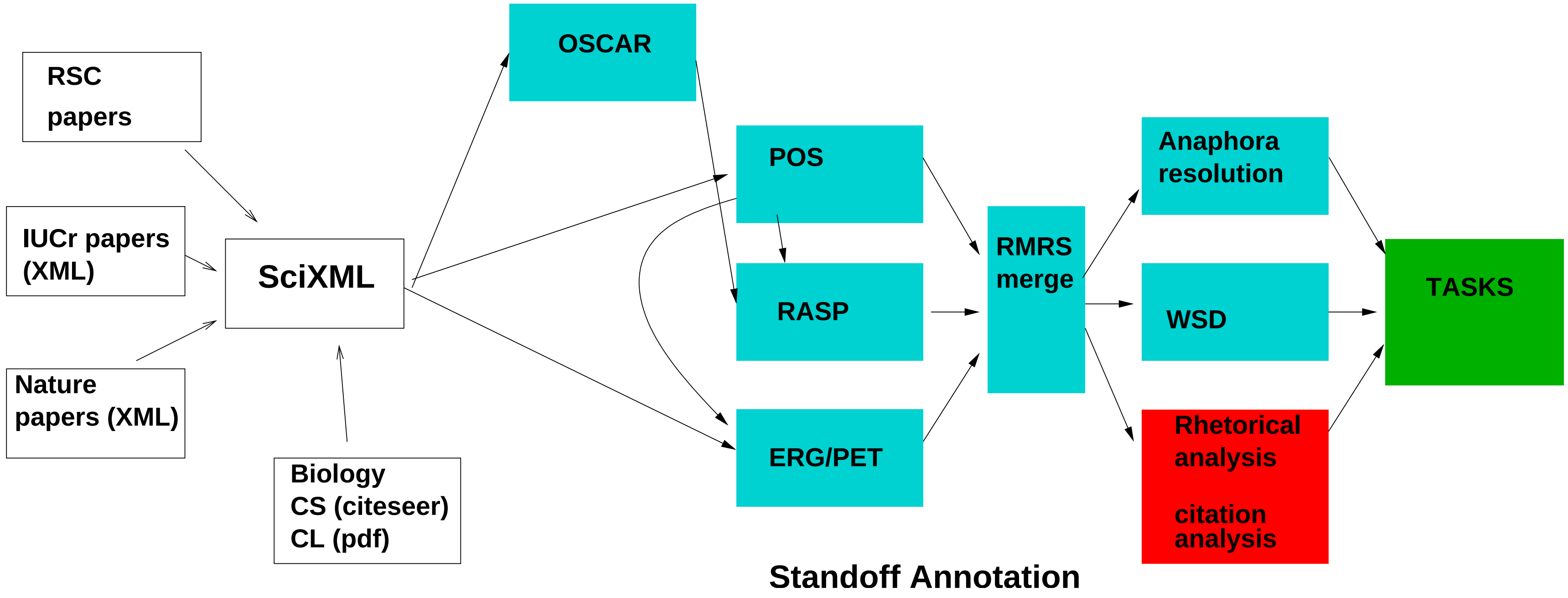
Project SciBorg: Knowledge Management for Chemists

- PIs Copestake, Teufel, Murray-Rust (Chemistry Dept Cambridge), Parker (CeSC)
- RAs CJ Rupp, Peter Corbett, Advaith Siddharthan
- Partners: Nature, Royal Society of Chemistry, International Union of Crystallography
- Aims:
  - Develop a NL markup language (RMRS) which acts as platform for IE. Link to semantic web languages.
  - Develop IE technology and ontologies for use by publishers, researchers, readers, vendors, and regulatory organisations
  - Model scientific argumentation and citation function for new ways of information access
  - Demonstrate applicability of this infrastructure in real-world eScience.

Project SciBorg: Overview



Project SciBorg: Rhetorical Analysis



RSC papers

IUCr papers (XML)

Nature papers (XML)

Biology
CS (citeseer)
CL (pdf)

SciXML

OSCAR

POS

RASP

ERG/PET

RMRS merge

Anaphora resolution

WSD

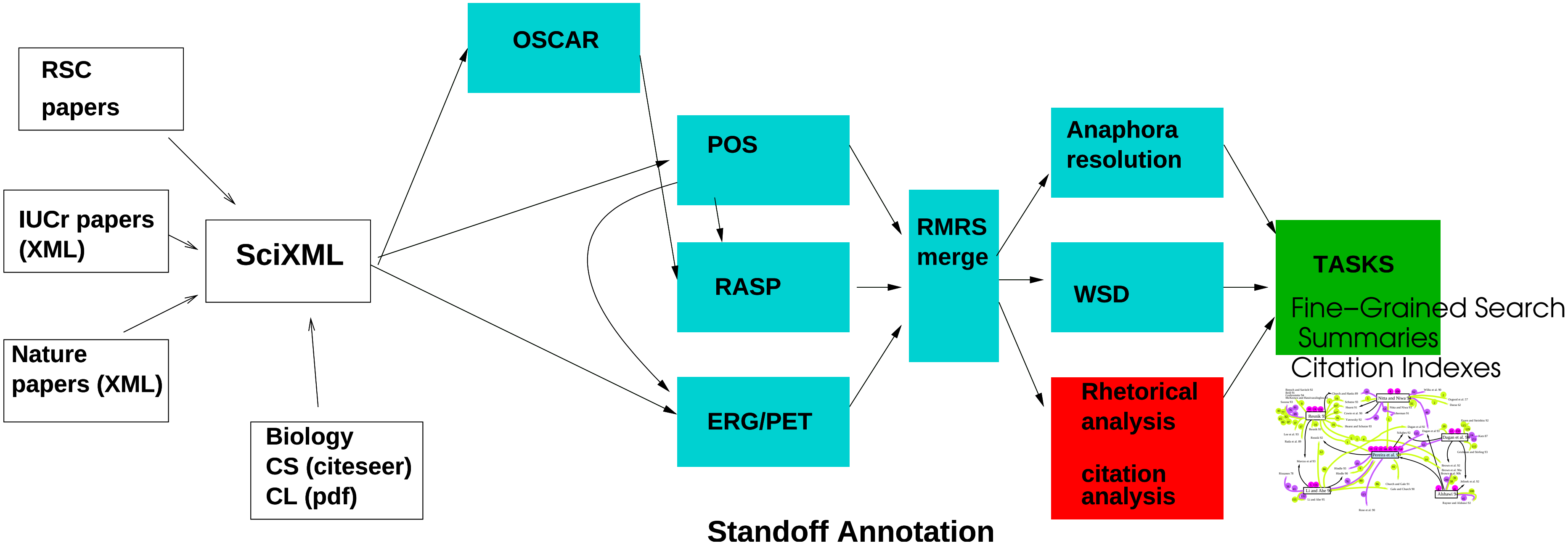
Rhetorical analysis

citation analysis

TASKS

Standoff Annotation

Project SciBorg: Novel Information Access



RSC papers

IUCr papers (XML)

Nature papers (XML)

Biology
CS (citeseer)
CL (pdf)

SciXML

OSCAR

POS

RASP

ERG/PET

RMRS merge

Anaphora resolution

WSD

Rhetorical analysis

citation analysis

TASKS

Fine–Grained Search
Summaries
Citation Indexes

Standoff Annotation

Application: Chemical Search

Search for papers describing the synthesis of Troeger’s base from anilines:

The synthesis of 2,8-dimethyl-6H, 12H-5, 11methanodibenzo[b,f]diazocine (Troeger’s base) from p-toluidine and of two other Troger’s base analogs from other aninlines

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

Even harder: search for papers describing synthesis of Troeger’s base which **don’t** involve anilines

Retrieve all papers `X: Goal(X,h), h:synthesis, result(h,<TB>), Source(h,y) & NOT(aniline(y))`

Applications: Organic syntheses and Ontologies

- *To a solution of aldimine$_1$ (1.5mmol) in THF (5ml) was added LDA (1ml, 1.6 M in THF) at 0° under argon, the resulting mixture was stirred for 2h, then was cooled to -78° . . .*

  → recipe expressed in CML extension formalism

- *. . . alkaloids and other complex polycyclic azacycles . . .*

  →
  ```
  <owl:class rdf:ID="Alkaloid"><rdfs:SubClassOf
  rdf:resource="#Azacycle"/></owl:class>
  ```

Application: Fine-Grained (Rhetorical) Search

Chemists search for descriptions of failed problem solving activity:

> *. . . suggested the possibility of exploiting this steroid for the generation of a chiral but non-C2-symmetrical macrocyclic barbiturate binding receptor. To achieve this goal, the corresponding methyl ester was reduced and tritylated to afford the monoprotected triol* **11** *(Scheme 3). Subsequent chain extension proceeded as previously described for the synthesis of the macrocycle* **3a** *affording the dicarboxylic acid* **13**. *Nevertheless, several attempts to promote the cyclization under high dilution conditions between the corresponding acid chloride of* **13** *with diamide* **10** *only led to trace amounts of the desired macrocycle* **14**. *Reversing the roles of the acylating agent, however, proved more rewarding.*

Another Application: Fine-Grained (Rhetorical) Search

Differences between compounds (in terms of properties, chemical structure, preparation or appliations):

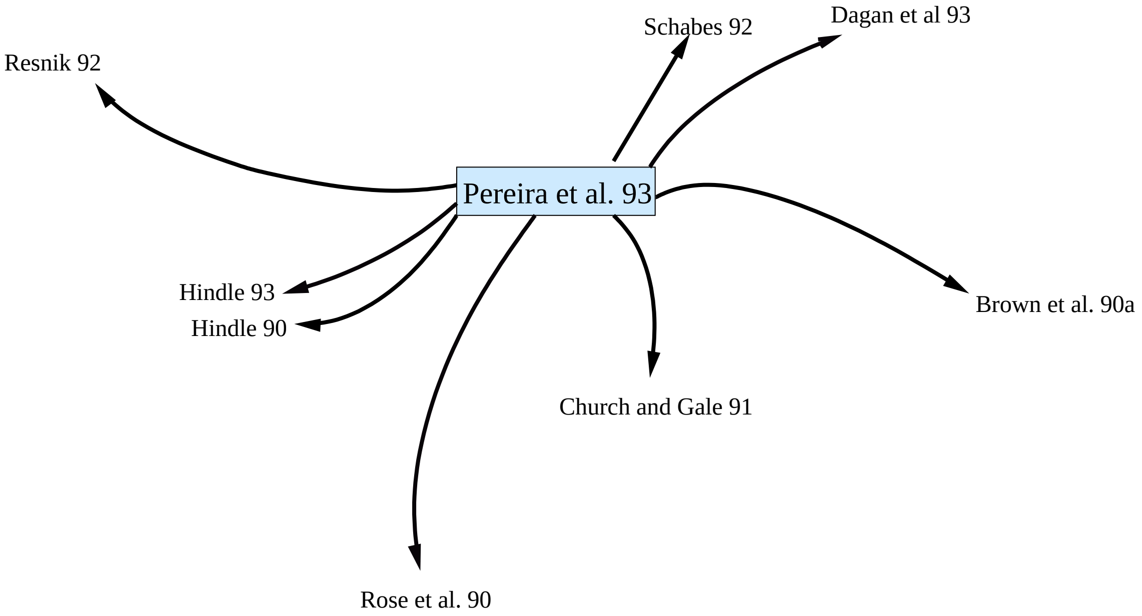
> *Most of the analogues have comparable antimalarial IC50 values to the naturally occurring endoperoxide artemisinin.*

> *Notably, the spiro-amides* **37–40** *have much lower potency than members of the dispiro series.*

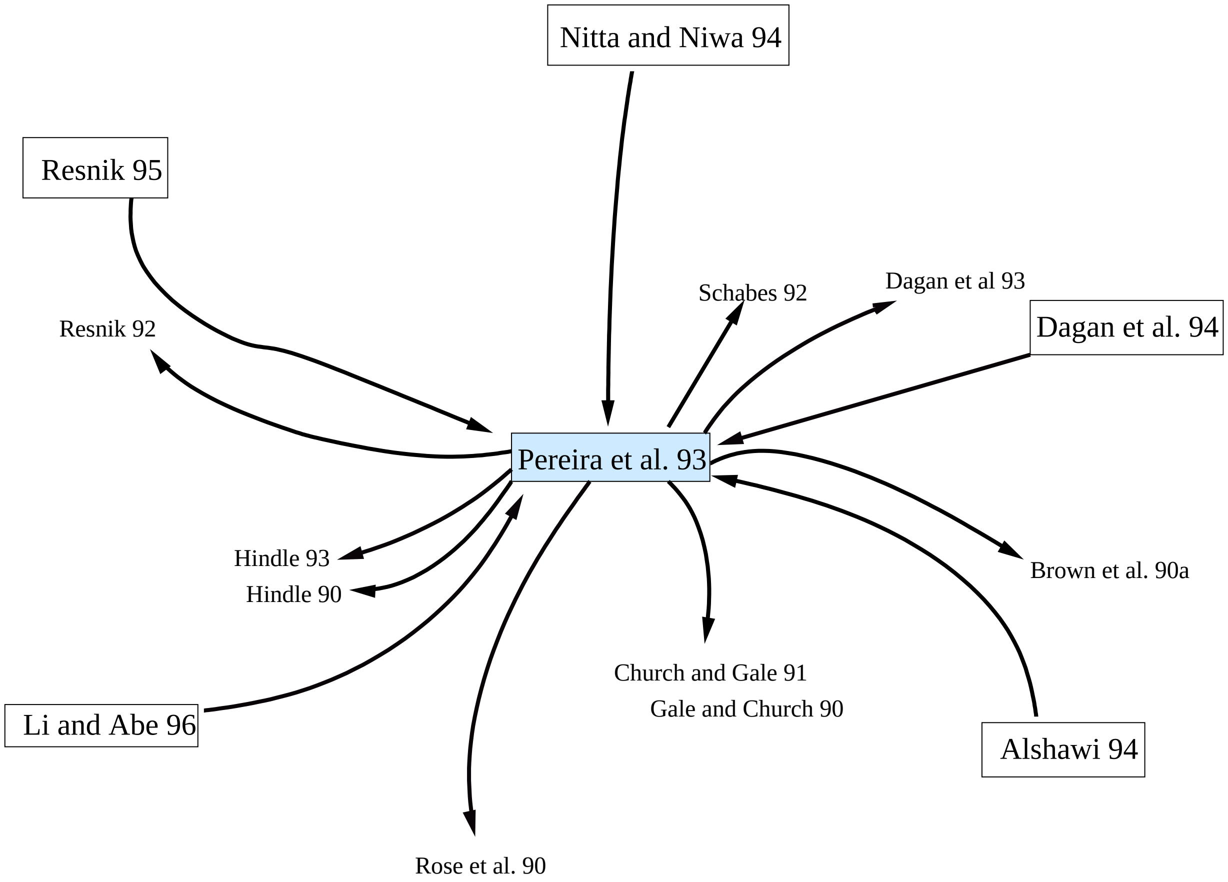
Find “strong claims” in the literature (Project FlyBase):

> *In contrast with previous hypotheses, compact plaques form before significant deposition of diffuse A beta, suggesting that different mechanisms are involved in the deposition of diffuse amyloid and the aggregation into plaques.*

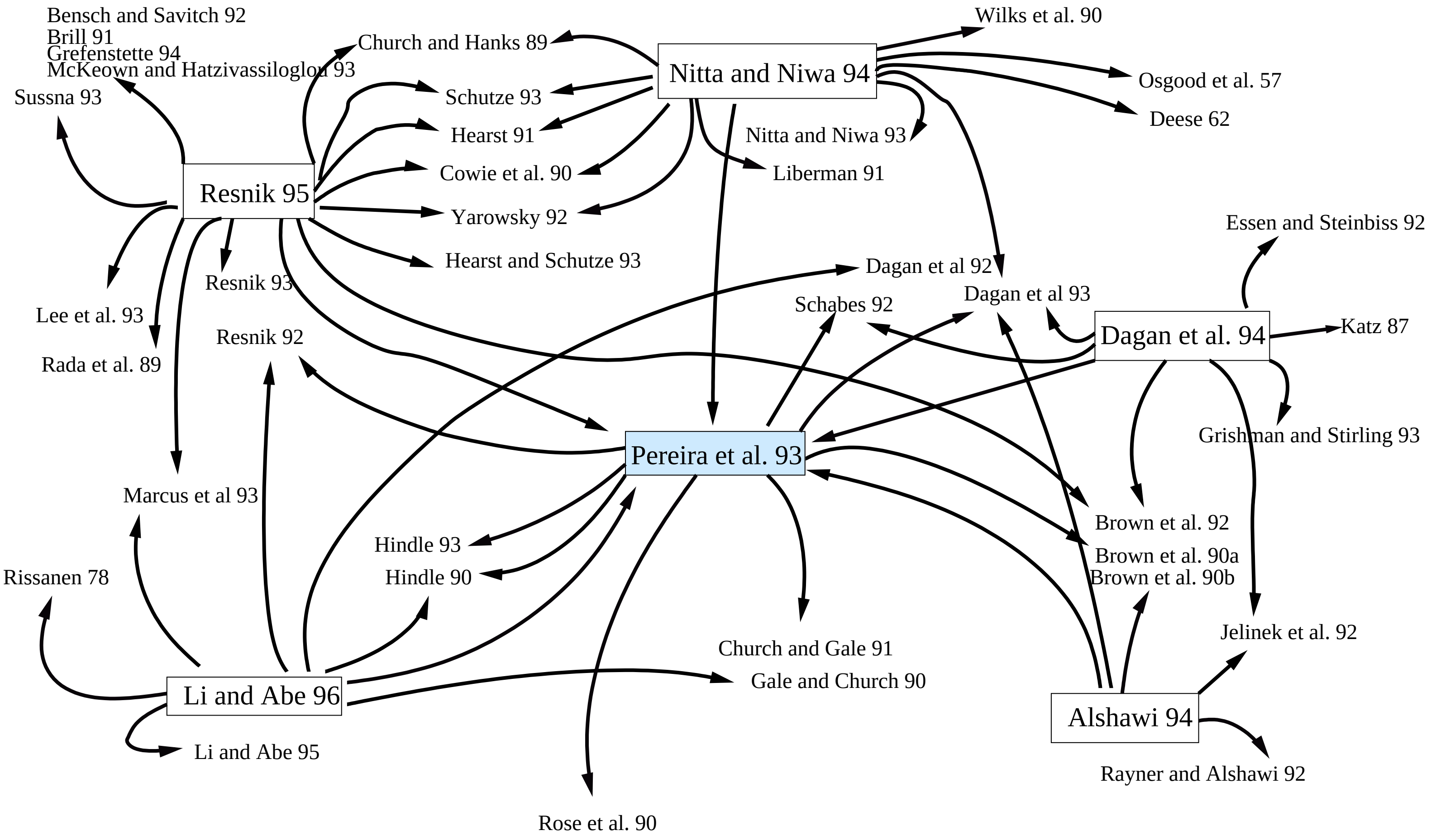
Another Application: Citation Map (project CitRAZ)



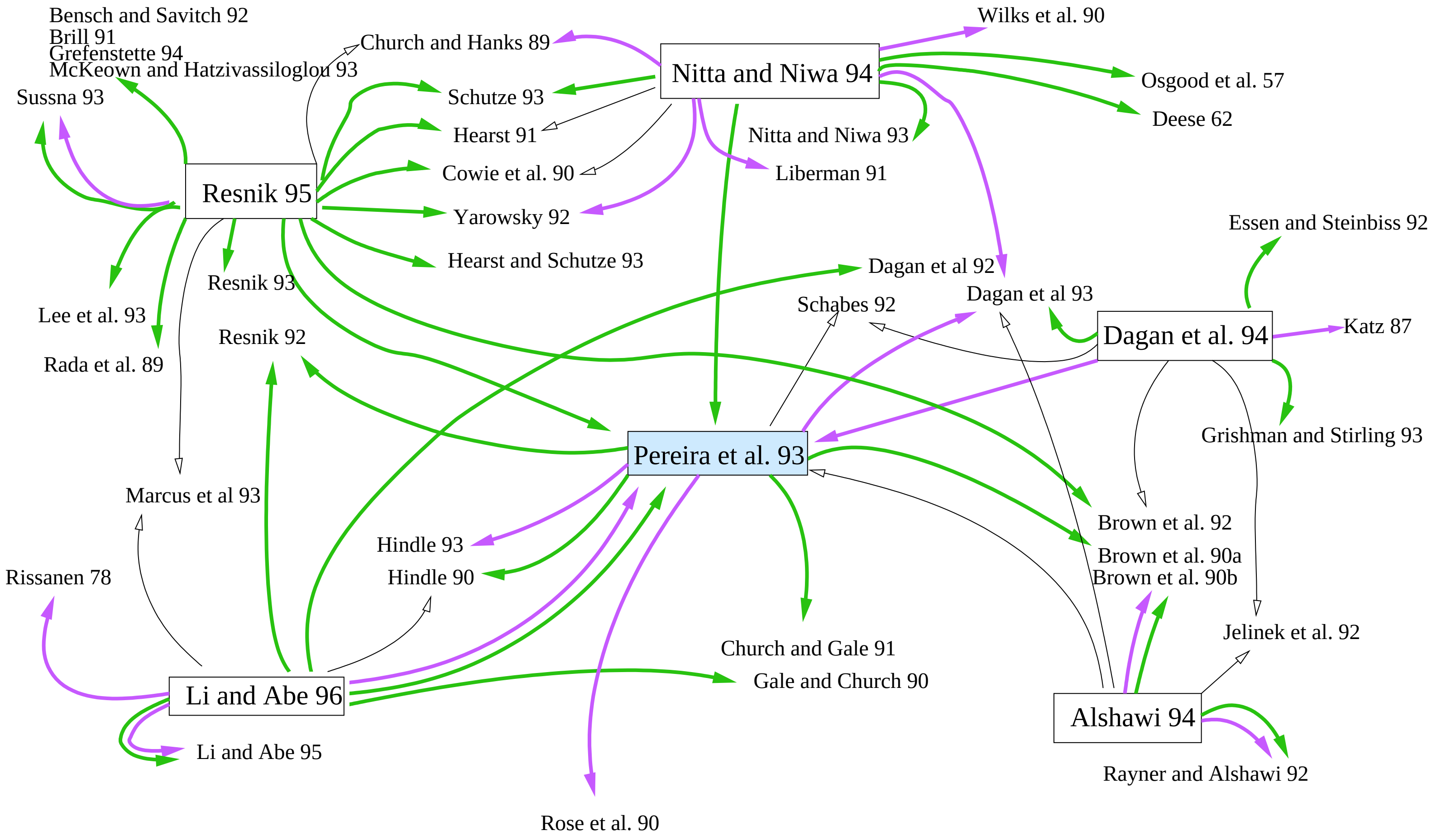
Another Application: Citation Map (project CitRAZ)



Another Application: Citation Map (project CitRAZ)

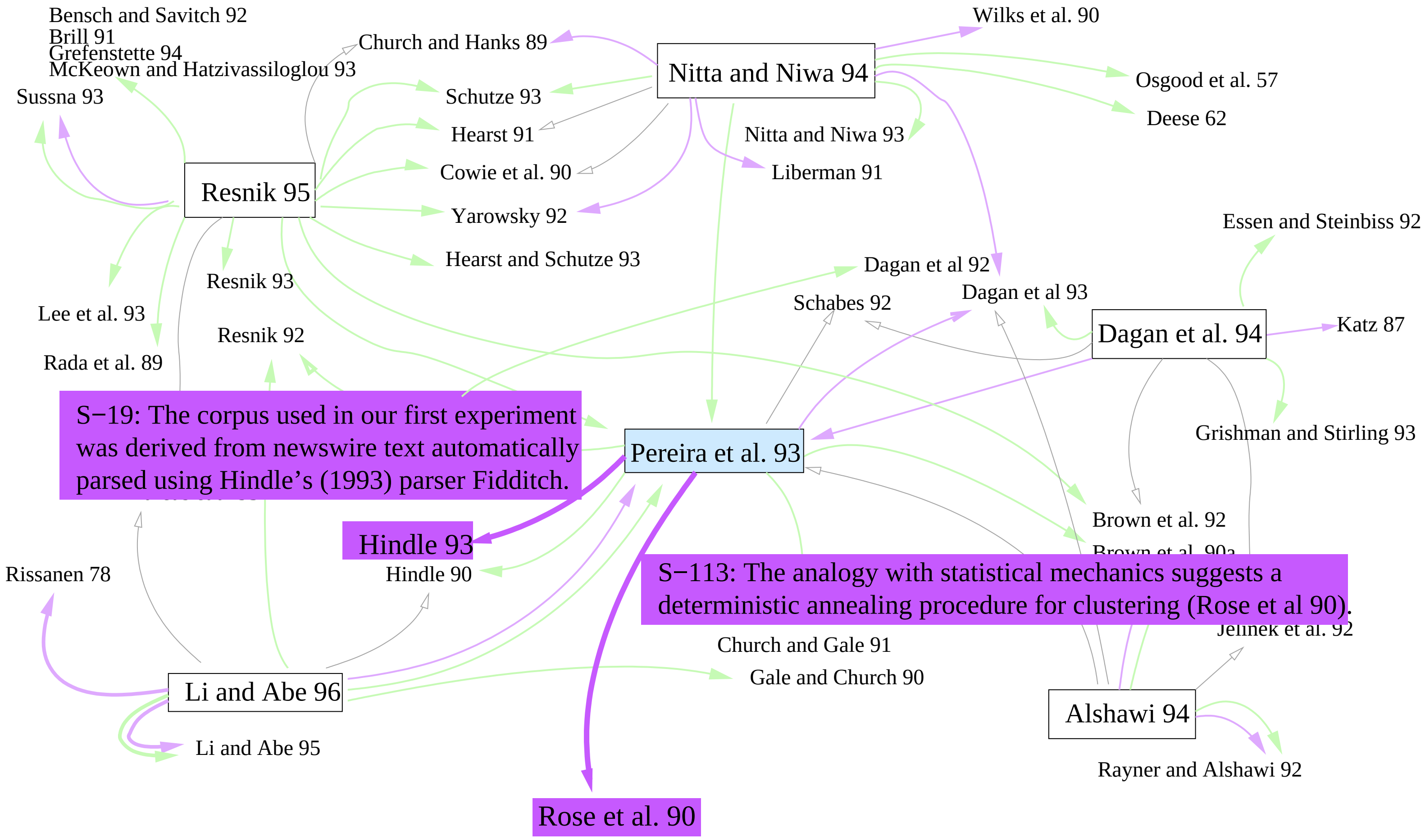


Another Application: Citation Map (project CitRAZ)



Bensch and Savitch 92
Brill 91
Grefenstette 94
McKeown and Hatzivassiloglou 93
Sussna 93
Church and Hanks 89
Nitta and Niwa 94
Wilks et al. 90
Osgood et al. 57
Deese 62
Schutze 93
Hearst 91
Nitta and Niwa 93
Cowie et al. 90
Liberman 91
Resnik 95
Yarowsky 92
Hearst and Schutze 93
Essen and Steinbiss 92
Dagan et al 92
Dagan et al 93
Resnik 93
Schabes 92
Lee et al. 93
Dagan et al. 94
Katz 87
Resnik 92
Rada et al. 89
Grishman and Stirling 93
Pereira et al. 93
Marcus et al 93
Brown et al. 92
Hindle 93
Brown et al. 90a
Brown et al. 90b
Rissanen 78
Hindle 90
Jelinek et al. 92
Church and Gale 91
Li and Abe 96
Gale and Church 90
Alshawi 94
Li and Abe 95
Rayner and Alshawi 92
Rose et al. 90

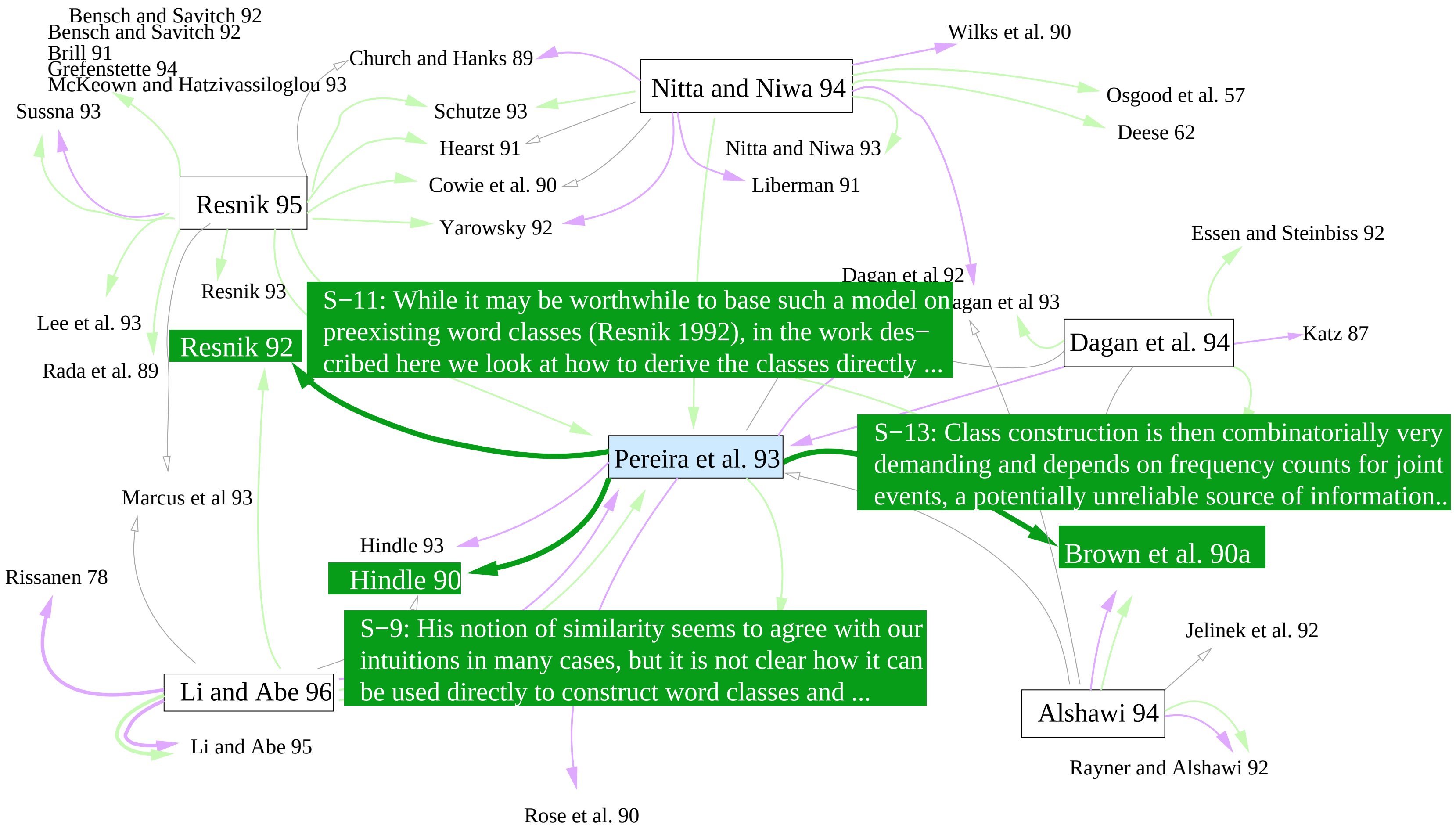
Another Application: Citation Map (project CitRAZ)



S−19: The corpus used in our first experiment was derived from newswire text automatically parsed using Hindle's (1993) parser Fidditch.

S−113: The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering (Rose et al 90).

Another Application: Citation Map (project CitRAZ)

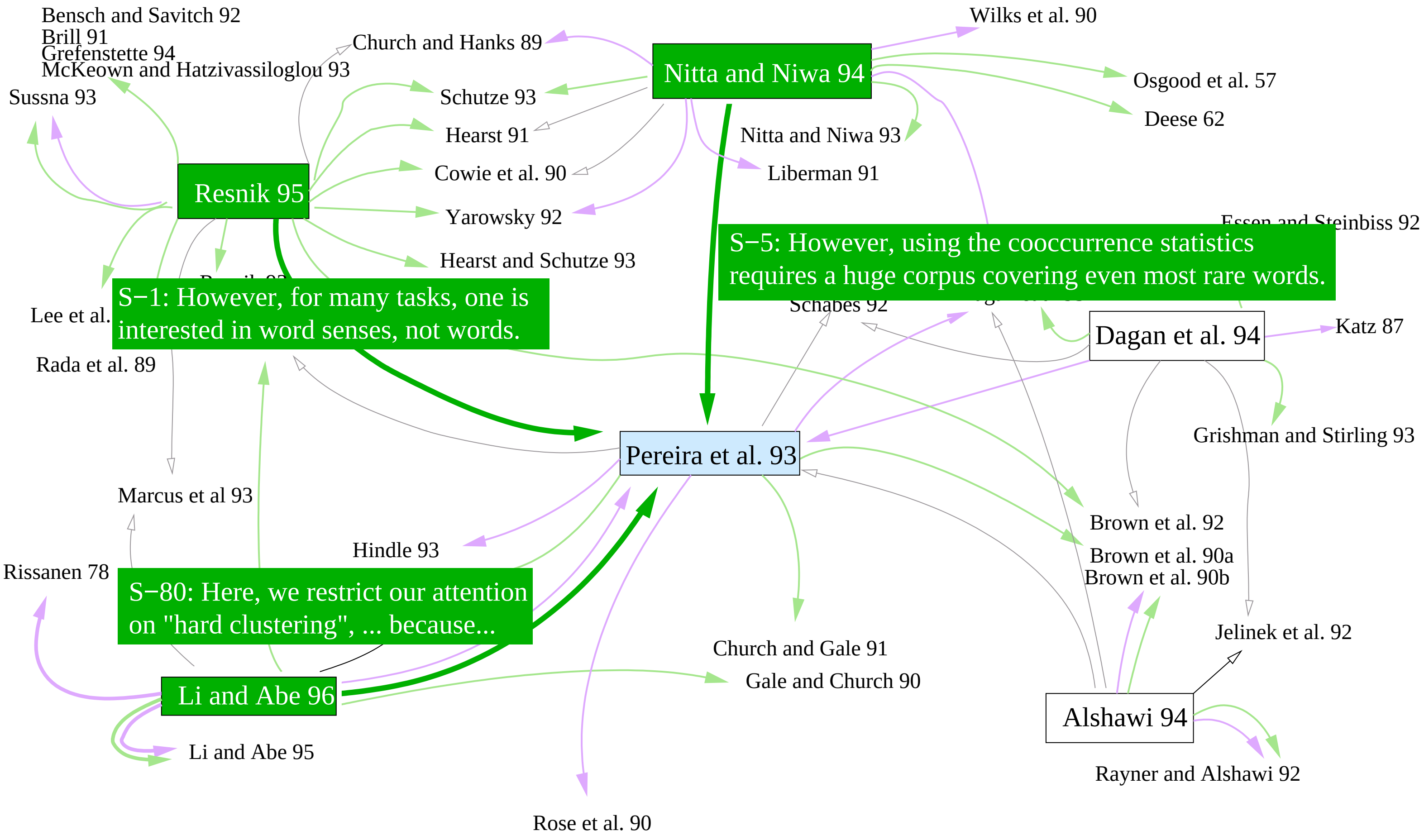


S−11: While it may be worthwhile to base such a model on preexisting word classes (Resnik 1992), in the work des−cribed here we look at how to derive the classes directly ...

S−13: Class construction is then combinatorially very demanding and depends on frequency counts for joint events, a potentially unreliable source of information..

S−9: His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and ...

Another Application: Citation Map (project CitRAZ)



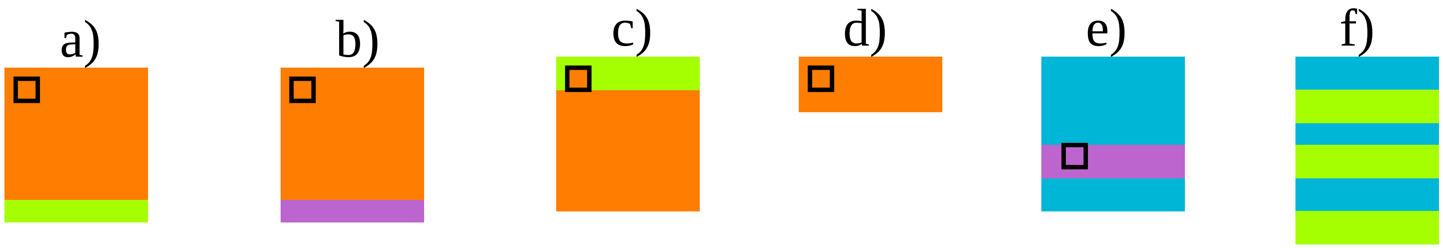
S−1: However, for many tasks, one is interested in word senses, not words.

S−5: However, using the cooccurrence statistics requires a huge corpus covering even most rare words.

S−80: Here, we restrict our attention on "hard clustering", ... because...

Citation Context

S−5 Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. S−6 For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. S−7 This requires a reasonable definition of verb similarity and a similarity estimation method. S−8 In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. S−9 His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.



Fact: 69% of the 600 CONTRAST sentences and 21% of the 246 BASIS sentences do not contain the citation!

Argumentative Zoning (Teufel and Moens, 2002)

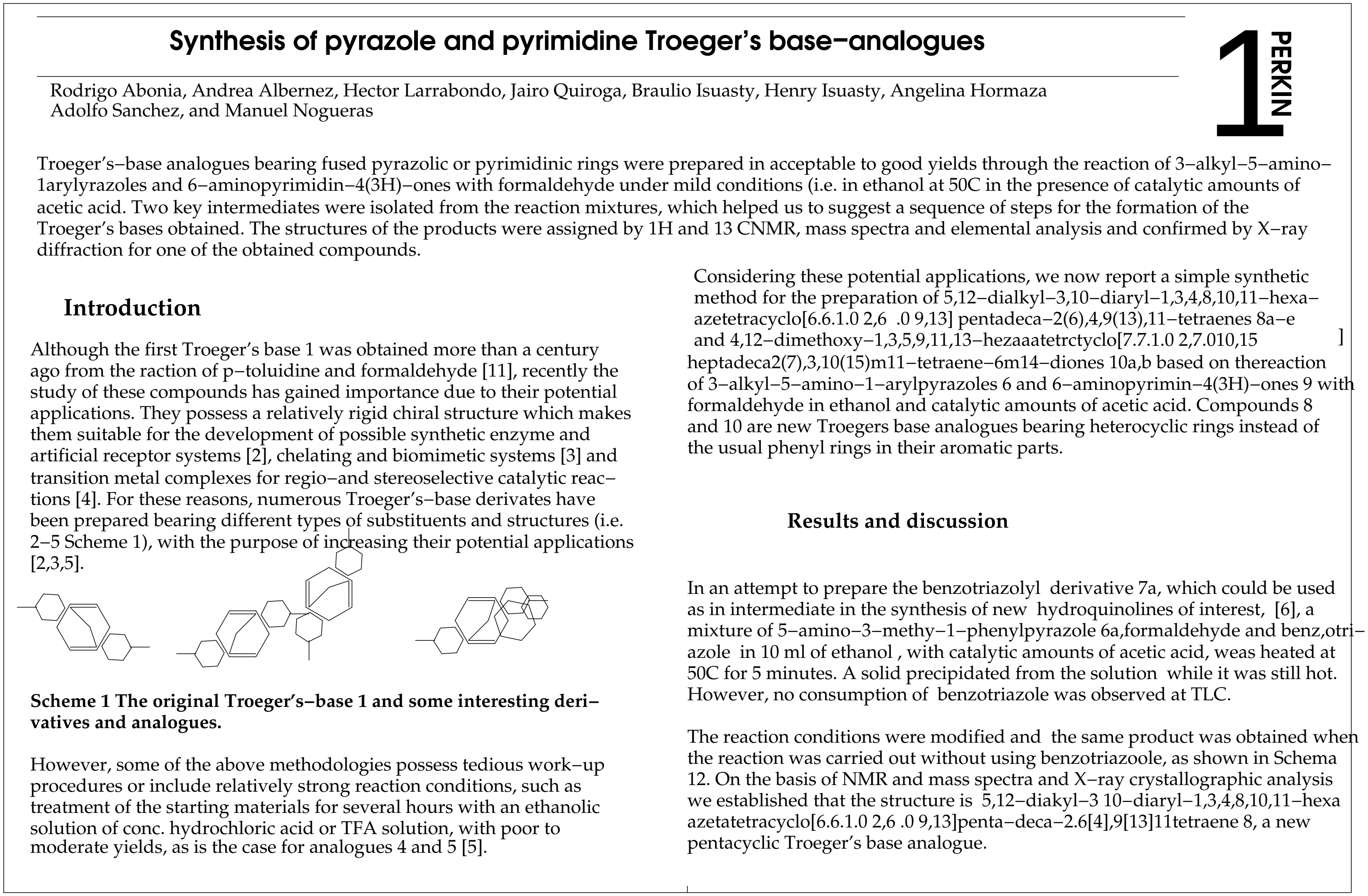
Method for robust analysis of rhetorical document structure

- Divide document into zones of same rhetorical status (Swales, 1990)
  - Segm. by Intellectual Ownership: Own – Other – General
  - Problem Solving Structure: prototypical statements
    - Problem-Solving Activity has failed/was successful
    - Problem is hard, solution is novel
  - Stance towards cited work plays a role in the argumentation

AIM	Statements of author's scientific aim
CONTRAST	Statements of difference with other work
BASIS	Statements of origins of ideas
OTHER	Neutral description of other work
BACKGROUND	Generally accepted statements in the field
TEXTUAL	Statements about external structure of article
OWN	All other statements about own work

- Uses supervised ML to simulate human annotation

Document Analysis in SciBorg



Synthesis of pyrazole and pyrimidine Troeger's base–analogues

Rodrigo Abonia, Andrea Albernez, Hector Larrabondo, Jairo Quiroga, Braulio Isuasty, Henry Isuasty, Angelina Hormaza
Adolfo Sanchez, and Manuel Nogueras

1 PERKIN

Troeger's–base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3–alkyl–5–amino–1arylyrazoles and 6–aminopyrimidin–4(3H)–ones with formaldehyde under mild conditions (i.e. in ethanol at 50C 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 Troeger's bases obtained. The structures of the products were assigned by 1H and 13 CNMR, mass spectra and elemental analysis and confirmed by X–ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base 1 was obtained more than a century ago from the raction of p–toluidine and formaldehyde [11], 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 Troeger's–base derivates have been prepared bearing different types of substituents and structures (i.e. 2–5 Scheme 1), with the purpose of increasing their potential applications [2,3,5].

Scheme 1 The original Troeger'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 [5].

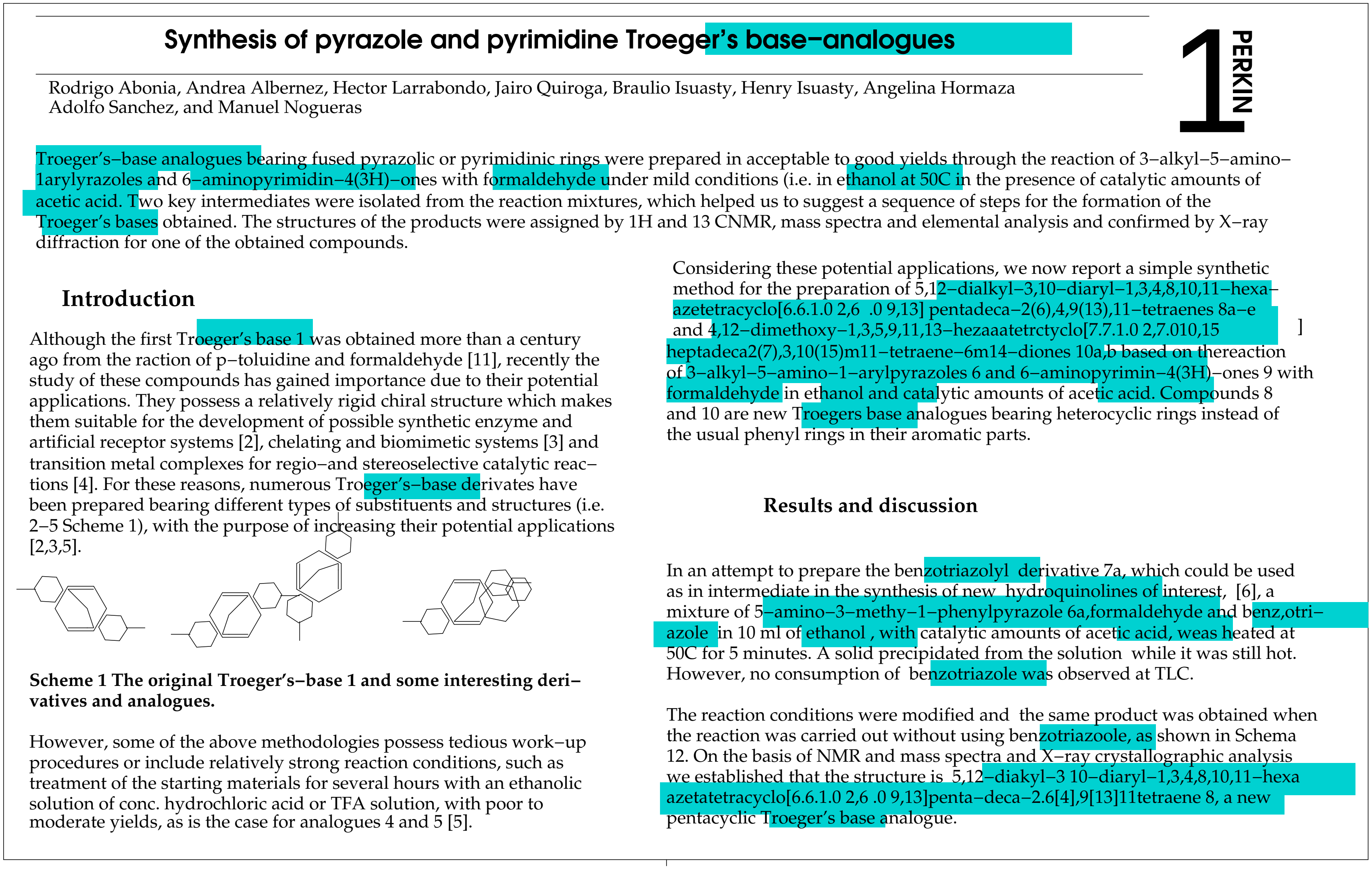
Considering these potential applications, we now report a simple synthetic method for the preparation of 5,12–dialkyl–3,10–diaryl–1,3,4,8,10,11–hexa–azetetracyclo[6.6.1.0 2,6 .0 9,13] pentadeca–2(6),4,9(13),11–tetraenes 8a–e and 4,12–dimethoxy–1,3,5,9,11,13–hezaaatetrctyclo[7.7.1.0 2,7.010,15] heptadeca2(7),3,10(15)m11–tetraene–6m14–diones 10a,b based on thereaction of 3–alkyl–5–amino–1–arylpyrazoles 6 and 6–aminopyrimin–4(3H)–ones 9 with formaldehyde in ethanol and catalytic amounts of acetic acid. Compounds 8 and 10 are new Troegers 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 in intermediate in the synthesis of new hydroquinolines of interest, [6], a mixture of 5–amino–3–methy–1–phenylpyrazole 6a,formaldehyde and benz,otri–azole in 10 ml of ethanol , with catalytic amounts of acetic acid, weas heated at 50C for 5 minutes. A solid precipidated from the solution while it was still hot. However, no consumption of benzotriazole was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazoole, as shown in Schema 12. On the basis of NMR and mass spectra and X–ray crystallographic analysis we established that the structure is 5,12–diakyl–3 10–diaryl–1,3,4,8,10,11–hexa azetatetracyclo[6.6.1.0 2,6 .0 9,13]penta–deca–2.6[4],9[13]11tetraene 8, a new pentacyclic Troeger's base analogue.

Document Analysis: OSCAR



Synthesis of pyrazole and pyrimidine Troeger's base–analogues

1 PERKIN

Rodrigo Abonia, Andrea Albernez, Hector Larrabondo, Jairo Quiroga, Braulio Isuasty, Henry Isuasty, Angelina Hormaza
Adolfo Sanchez, and Manuel Nogueras

Troeger's–base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3–alkyl–5–amino–1arylyrazoles and 6–aminopyrimidin–4(3H)–ones with formaldehyde under mild conditions (i.e. in ethanol at 50C 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 Troeger's bases obtained. The structures of the products were assigned by 1H and 13 CNMR, mass spectra and elemental analysis and confirmed by X–ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base 1 was obtained more than a century ago from the raction of p–toluidine and formaldehyde [11], 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 Troeger's–base derivates have been prepared bearing different types of substituents and structures (i.e. 2–5 Scheme 1), with the purpose of increasing their potential applications [2,3,5].

Scheme 1 The original Troeger'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 [5].

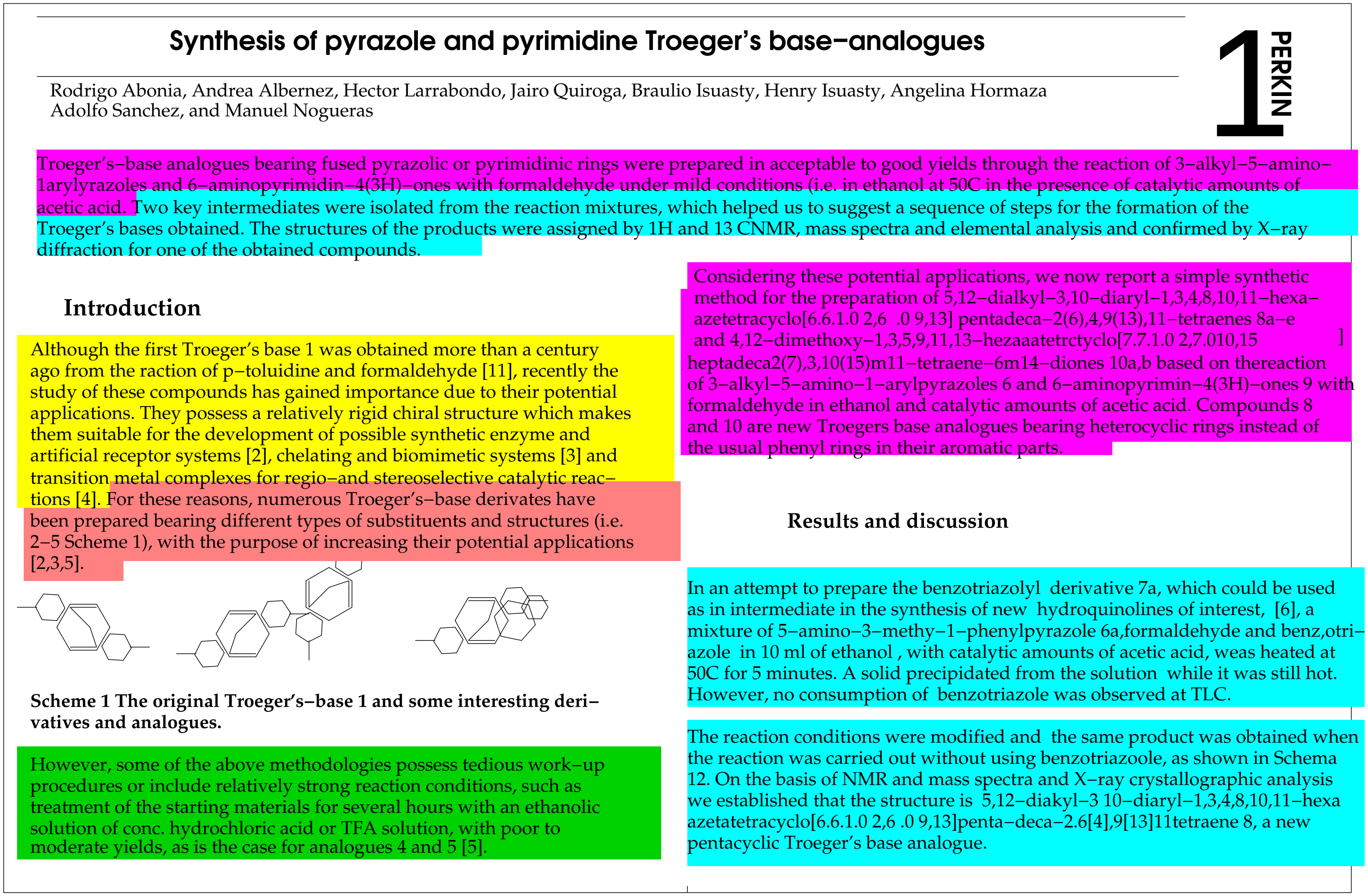
Considering these potential applications, we now report a simple synthetic method for the preparation of 5,12–dialkyl–3,10–diaryl–1,3,4,8,10,11–hexaazetetracyclo[6.6.1.0 2,6 .0 9,13] pentadeca–2(6),4,9(13),11–tetraenes 8a–e and 4,12–dimethoxy–1,3,5,9,11,13–hezaaatetrctyclo[7.7.1.0 2,7.010,15] heptadeca2(7),3,10(15)m11–tetraene–6m14–diones 10a,b based on thereaction of 3–alkyl–5–amino–1–arylpyrazoles 6 and 6–aminopyrimin–4(3H)–ones 9 with formaldehyde in ethanol and catalytic amounts of acetic acid. Compounds 8 and 10 are new Troegers 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 in intermediate in the synthesis of new hydroquinolines of interest, [6], a mixture of 5–amino–3–methy–1–phenylpyrazole 6a,formaldehyde and benz,otriazole in 10 ml of ethanol , with catalytic amounts of acetic acid, weas heated at 50C for 5 minutes. A solid precipidated from the solution while it was still hot. However, no consumption of benzotriazole was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazoole, as shown in Schema 12. On the basis of NMR and mass spectra and X–ray crystallographic analysis we established that the structure is 5,12–diakyl–3 10–diaryl–1,3,4,8,10,11–hexaazetatetracyclo[6.6.1.0 2,6 .0 9,13]penta–deca–2.6[4],9[13]11tetraene 8, a new pentacyclic Troeger's base analogue.

Document Analysis: Argumentative Zoning



Synthesis of pyrazole and pyrimidine Troeger's base–analogues

1 PERKIN

Rodrigo Abonia, Andrea Albernez, Hector Larrabondo, Jairo Quiroga, Braulio Isuasty, Henry Isuasty, Angelina Hormaza
Adolfo Sanchez, and Manuel Nogueras

Troeger's–base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3–alkyl–5–amino–1arylyrazoles and 6–aminopyrimidin–4(3H)–ones with formaldehyde under mild conditions (i.e. in ethanol at 50C 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 Troeger's bases obtained. The structures of the products were assigned by 1H and 13 CNMR, mass spectra and elemental analysis and confirmed by X–ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base 1 was obtained more than a century ago from the raction of p–toluidine and formaldehyde [11], 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 Troeger's–base derivates have been prepared bearing different types of substituents and structures (i.e. 2–5 Scheme 1), with the purpose of increasing their potential applications [2,3,5].

Scheme 1 The original Troeger'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 [5].

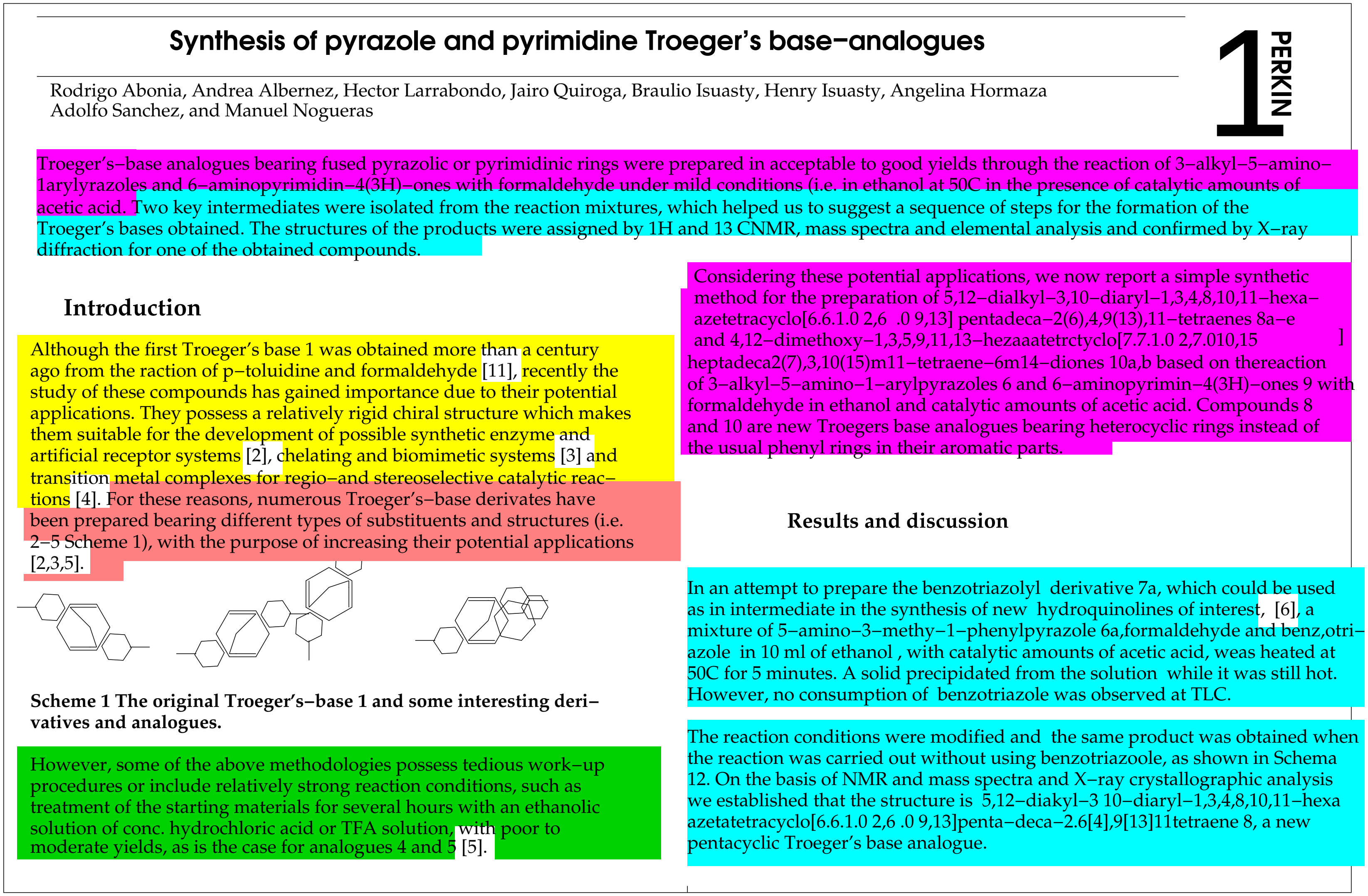
Considering these potential applications, we now report a simple synthetic method for the preparation of 5,12–dialkyl–3,10–diaryl–1,3,4,8,10,11–hexaazetetracyclo[6.6.1.0 2,6 .0 9,13] pentadeca–2(6),4,9(13),11–tetraenes 8a–e and 4,12–dimethoxy–1,3,5,9,11,13–hezaaatetrctyclo[7.7.1.0 2,7.010,15] heptadeca2(7),3,10(15)m11–tetraene–6m14–diones 10a,b based on thereaction of 3–alkyl–5–amino–1–arylpyrazoles 6 and 6–aminopyrimin–4(3H)–ones 9 with formaldehyde in ethanol and catalytic amounts of acetic acid. Compounds 8 and 10 are new Troegers 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 in intermediate in the synthesis of new hydroquinolines of interest, [6], a mixture of 5–amino–3–methy–1–phenylpyrazole 6a,formaldehyde and benz,otriazole in 10 ml of ethanol , with catalytic amounts of acetic acid, weas heated at 50C for 5 minutes. A solid precipidated from the solution while it was still hot. However, no consumption of benzotriazole was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazoole, as shown in Schema 12. On the basis of NMR and mass spectra and X–ray crystallographic analysis we established that the structure is 5,12–diakyl–3 10–diaryl–1,3,4,8,10,11–hexa azetatetracyclo[6.6.1.0 2,6 .0 9,13]penta–deca–2.6[4],9[13]11tetraene 8, a new pentacyclic Troeger's base analogue.

Document Analysis: Citation Function Classification



Synthesis of pyrazole and pyrimidine Troeger's base–analogues

1 PERKIN

Rodrigo Abonia, Andrea Albernez, Hector Larrabondo, Jairo Quiroga, Braulio Isuasty, Henry Isuasty, Angelina Hormaza
Adolfo Sanchez, and Manuel Nogueras

Troeger's–base analogues bearing fused pyrazolic or pyrimidinic rings were prepared in acceptable to good yields through the reaction of 3–alkyl–5–amino–1arylyrazoles and 6–aminopyrimidin–4(3H)–ones with formaldehyde under mild conditions (i.e. in ethanol at 50C 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 Troeger's bases obtained. The structures of the products were assigned by 1H and 13 CNMR, mass spectra and elemental analysis and confirmed by X–ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base 1 was obtained more than a century ago from the raction of p–toluidine and formaldehyde [11], 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 Troeger's–base derivates have been prepared bearing different types of substituents and structures (i.e. 2–5 Scheme 1), with the purpose of increasing their potential applications [2,3,5].

Scheme 1 The original Troeger'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 [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–hexa–azetetracyclo[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–hezaaatetrctyclo[7.7.1.0 2,7.010,15] heptadeca2(7),3,10(15)m11–tetraene–6m14–diones 10a,b based on thereaction of 3–alkyl–5–amino–1–arylpyrazoles 6 and 6–aminopyrimin–4(3H)–ones 9 with formaldehyde in ethanol and catalytic amounts of acetic acid. Compounds 8 and 10 are new Troegers 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 in intermediate in the synthesis of new hydroquinolines of interest, [6], a mixture of 5–amino–3–methy–1–phenylpyrazole 6a,formaldehyde and benz,otri–azole in 10 ml of ethanol , with catalytic amounts of acetic acid, weas heated at 50C for 5 minutes. A solid precipidated from the solution while it was still hot. However, no consumption of benzotriazole was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzotriazoole, as shown in Schema 12. On the basis of NMR and mass spectra and X–ray crystallographic analysis we established that the structure is 5,12–diakyl–3 10–diaryl–1,3,4,8,10,11–hexa azetatetracyclo[6.6.1.0 2,6 .0 9,13]penta–deca–2.6[4],9[13]11tetraene 8, a new pentacyclic Troeger's base analogue.

Citations and Argumentation

Sentiment analysis around citations:

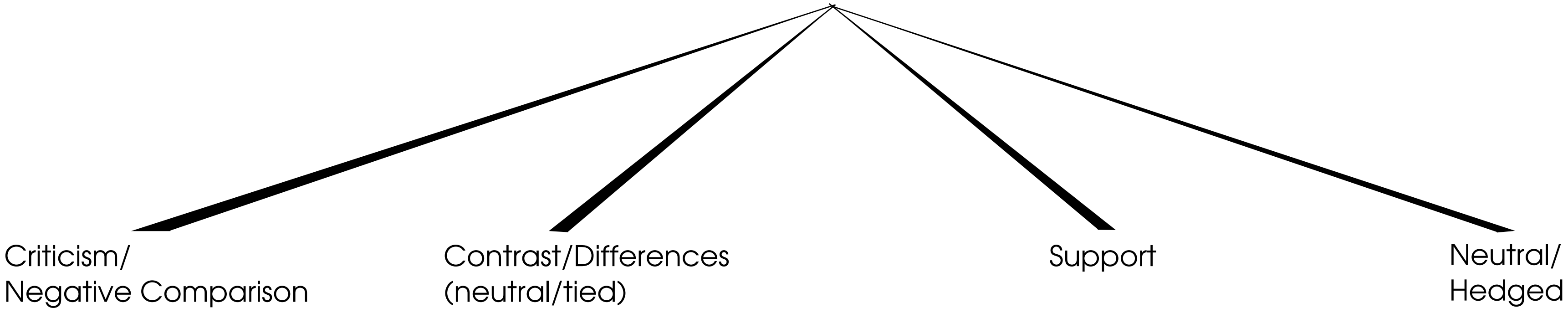
> *For these reasons numerous Tröger's base derivaties have been prepared . . . [2,3,5]. However, some of the above methodologies possess tedious work-up procedures or include relatively strong reaction conditions . . . with poor to moderate yields, as is the case for analogues* **4** *and* **5***.*

→ **Criticised** approach; typically in motivation

> *Conversion of the diol to the dicarboxylic acid* **8a** *was achieved by oxidation to the dialdezyde using the Dess-Martin periodinane*[13] *. . .*

→ **Used** approach; typically in description of own work

The CFC annotation scheme (CitRAZ)



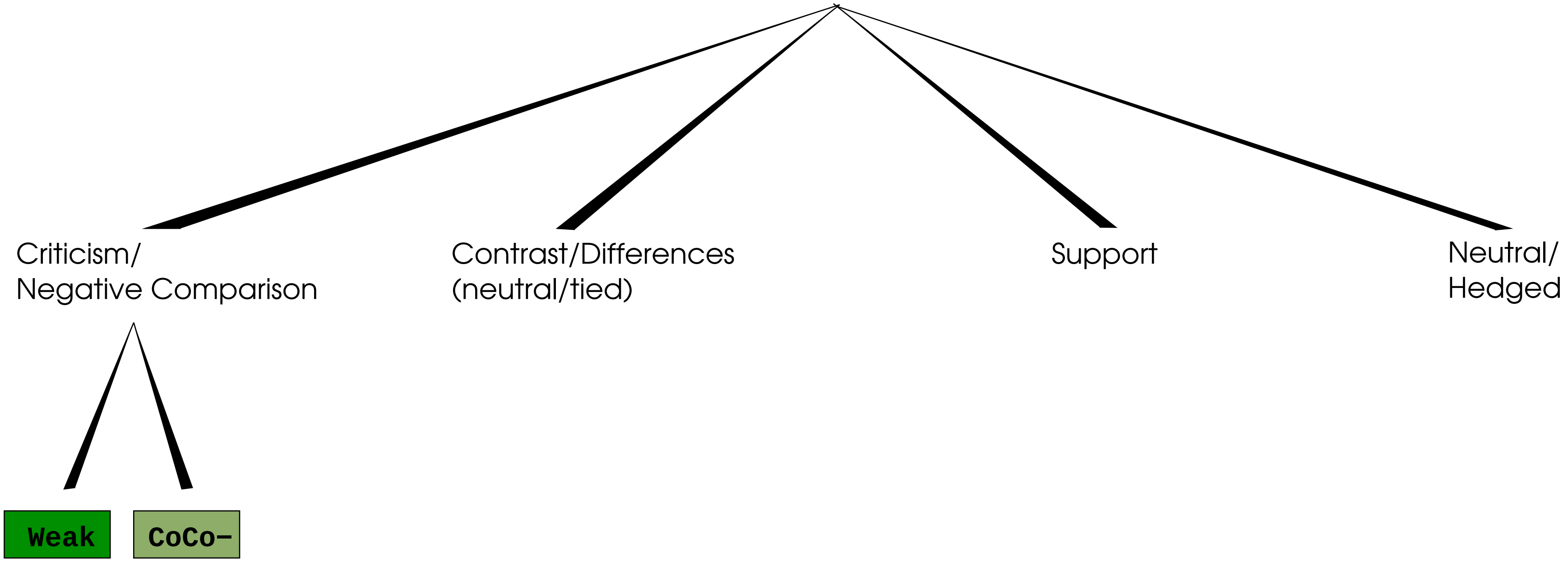
Criticism/
Negative Comparison

Contrast/Differences
(neutral/tied)

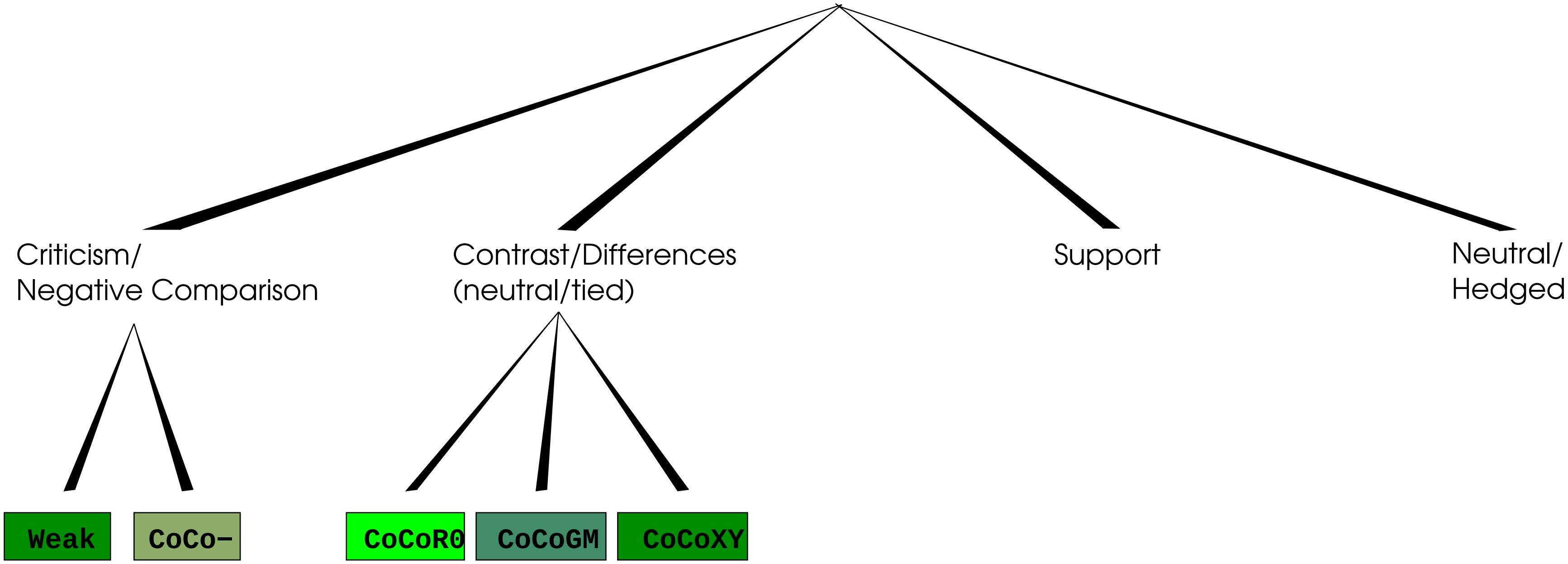
Support

Neutral/
Hedged

The CFC annotation scheme (CitRAZ)



The CFC annotation scheme (CitRAZ)



Criticism/
Negative Comparison

Contrast/Differences
(neutral/tied)

Support

Neutral/
Hedged

Weak

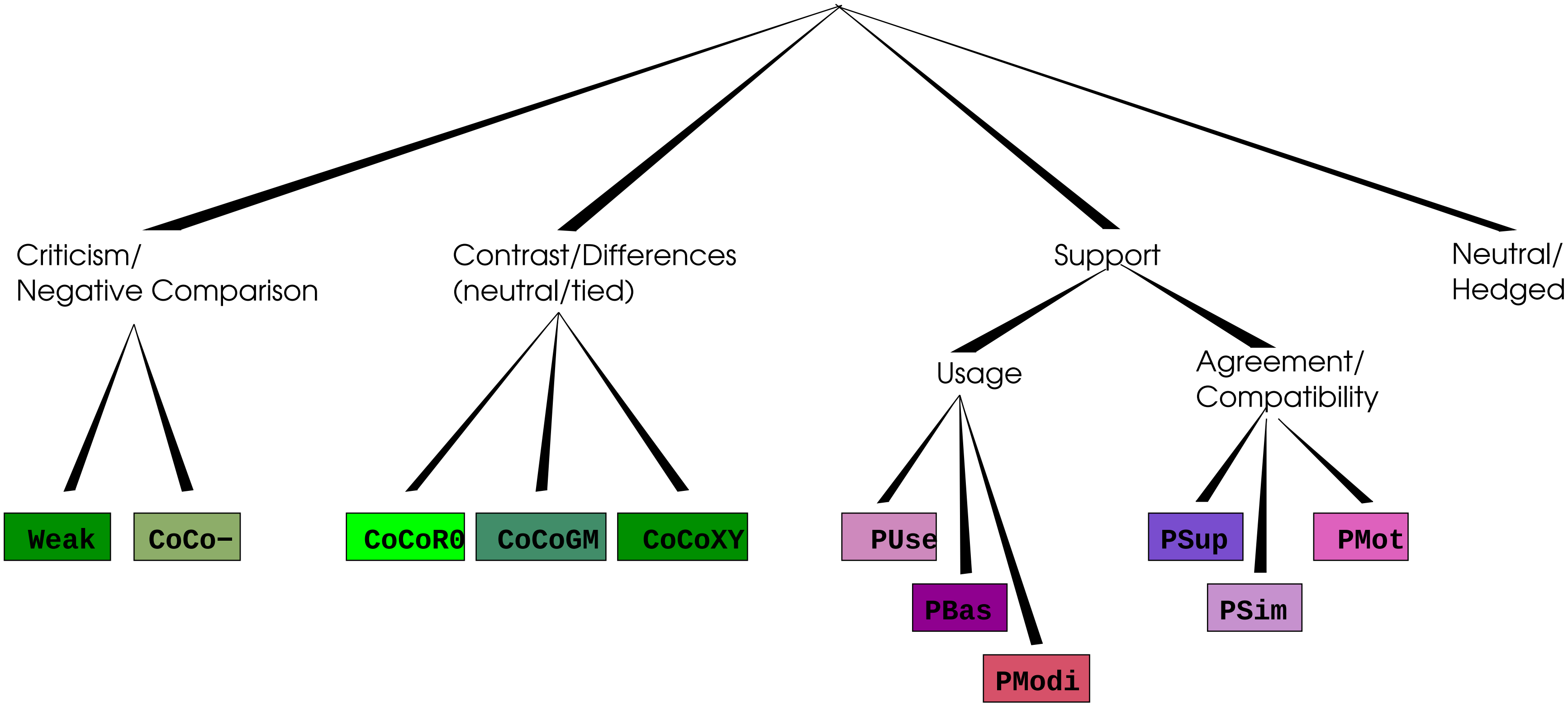
CoCo-

CoCoR0

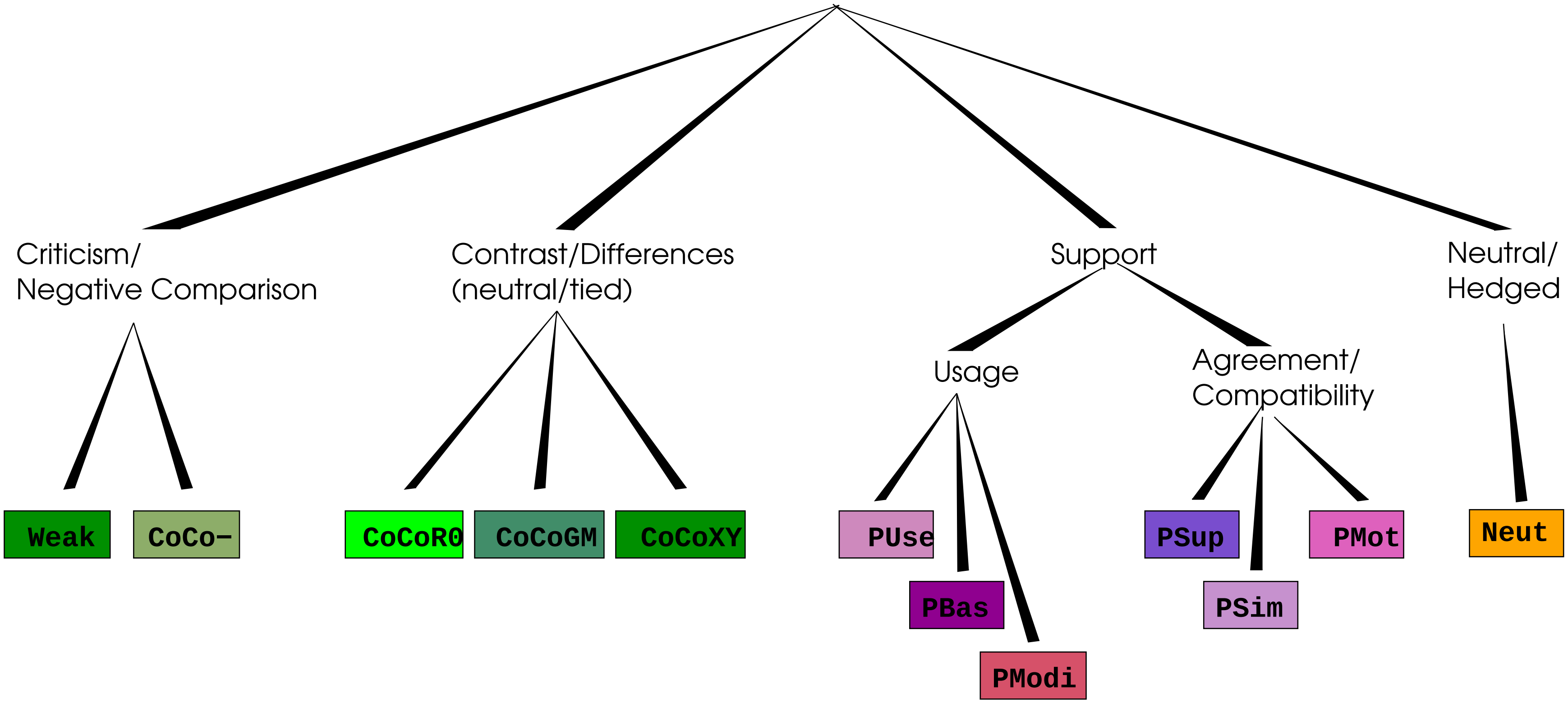
CoCoGM

CoCoXY

The CFC annotation scheme (CitRAZ)



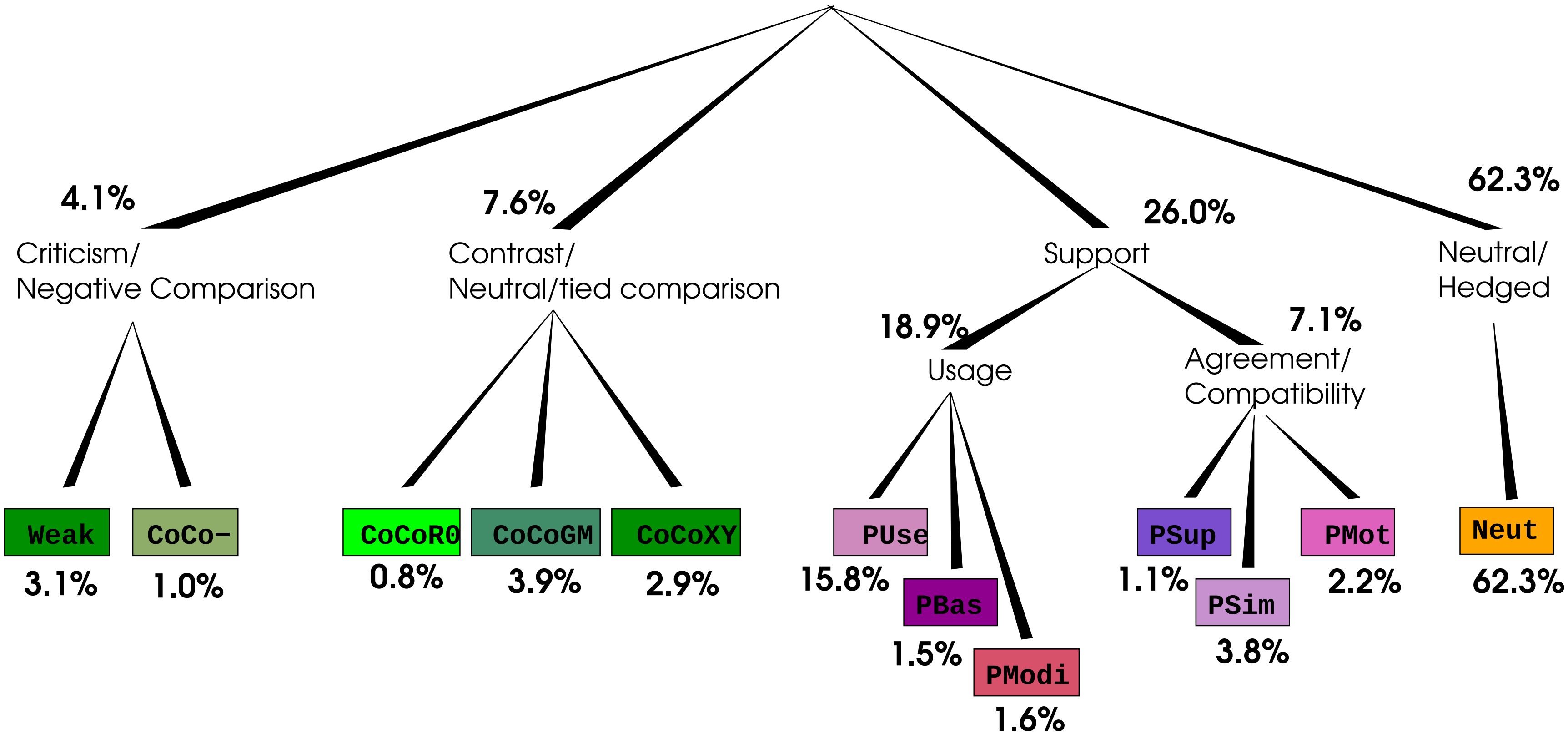
The CFC annotation scheme (CitRAZ)



Human Agreement (Teufel, Siddharthan, Tidhar 2006a)

- 3 task-trained annotators, 26 unseen articles, 548 citations
- Guidelines with over 120 rules
- Inter-annotator K=.72 (n=12;N=548;k=3)

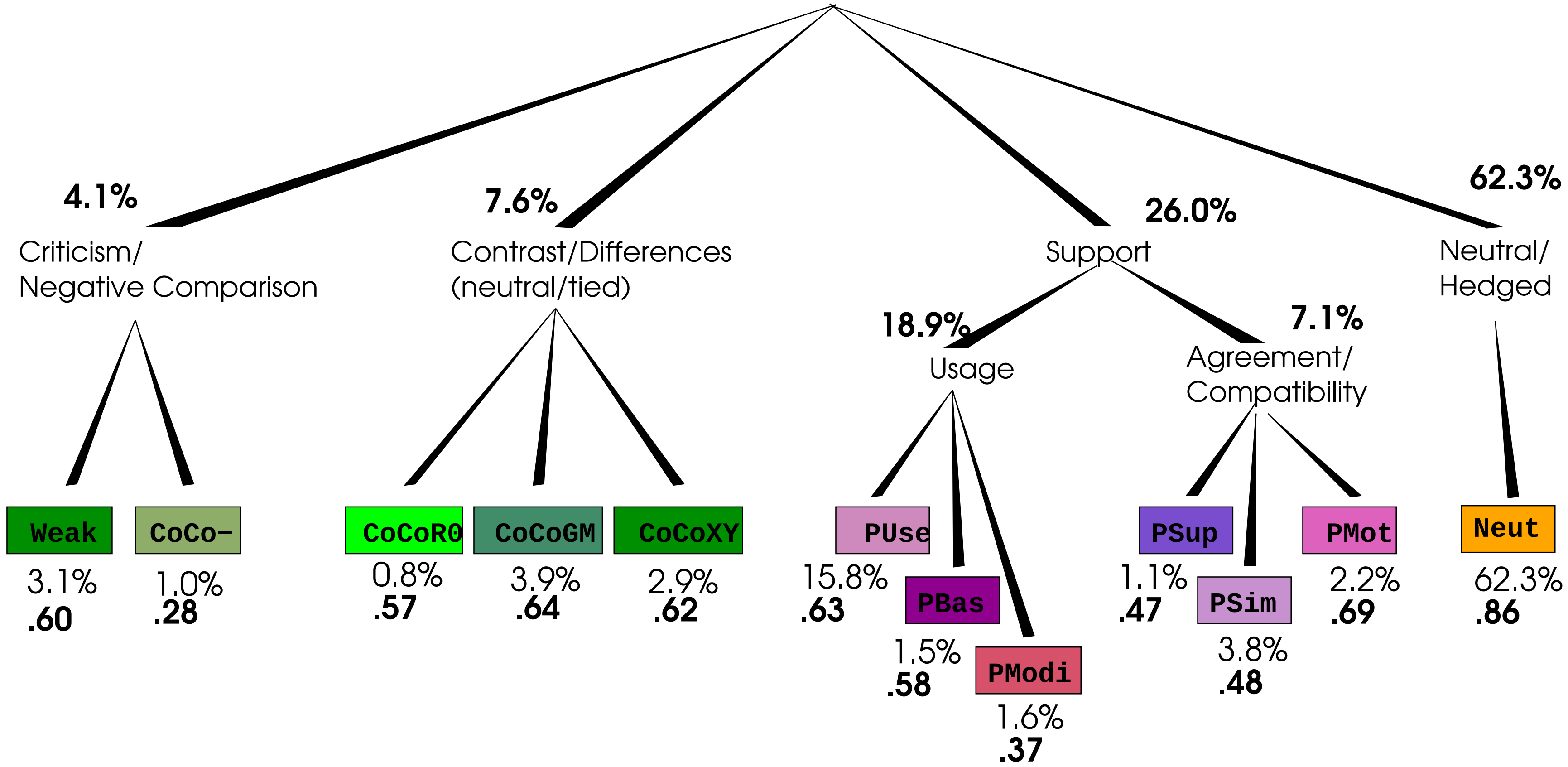
The CFC annotation scheme (CitRAZ)



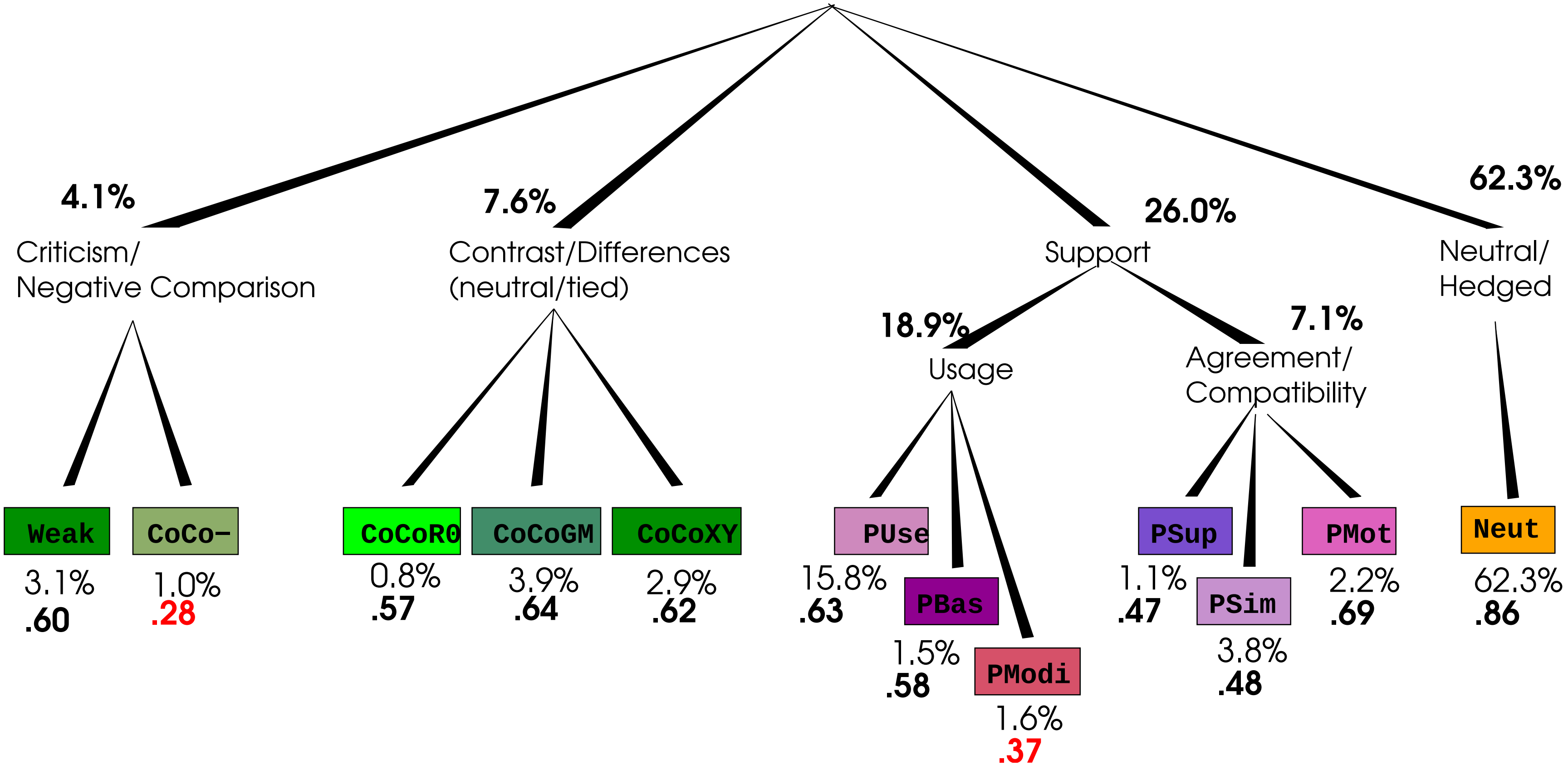
Features used for supervised ML

Type	Name	Feature description	Values
Absolute Location	Loc	Position of sentence in relation to 10 segments	10
Explicit Structure	Section Struct	Relative and absolute position of sentence within section (e.g., first sentence in section or somewhere in second third)	7
	Para Struct	Relative position of sentence within a paragraph	3
	Headline	Type of headline of current section	16
Sentence length	Length	Is the sentence longer than a certain threshold, measured in words?	2
Content Features	Title	Does the sentence contain words also occurring in the title or headlines?	2
	TF*IDF	Does the sentence contain “significant terms” as determined by the *TF*IDF* measure?	2
Verb Syntax	Voice	Voice (of first finite verb in sentence)	3
	Tense	Tense (of first finite verb in sentence)	10
	Modal	Is the first finite verb modified by modal auxiliary?	3
Citations	Cit	Citation or author name present? Self citation? Location of citation?	10
History	History	Most probable previous category	8
Meta-discourse	Formulaic	Type of formulaic expression occurring in sentence	28
	Agent	Type of Agent	10
	Action	Type of Action, with or without Negation	28

Classification results (Teufel et al. 2006b)



Classification results (Teufel et al. 2006b)



Results: Higher Levels of Hierarchy

- 4 Categ.: `Criticism` `Contrast` `Support` `Neutral`
- K =.59 (n=4, N=2829, k=2); P(A)=.79; Macro-F=.68
- Human agreement at K=.76 (P(A)=0.88, P(E)=0.46)

	`Criticism`	`Contrast`	`Support`	`Neutral`
Distr.	4.1%	7.6%	26.0%	62.3%
P	.80	.77	.75	.81
R	.49	.52	.65	.90
F	.61	.62	.70	.86

- All P above .75
- F around .7 (`Support`) and .6 (`Criticism`, `Contrast`)
- Effect of a) training material and b) “meek” citations

Projects SciBorg and CitRAZ: summary

- CitRAZ: Citation Maps, CL, only POS-level analysis
- SciBorg: Semantic analysis of each sentence performed
- Both: Discourse analysis: citations, general rhetorical status of sentence
- This allows for new forms of information access
  - fine-grained searches
  - citation maps
  - multi-document summaries
- SciBorg platform independent of discipline (Genetics, CS, Computational Linguistics)

AZ and CFC summary and outlook

- Improvement of AZ feature detection step by inclusion of a parser and anaphora resolver
- Application of AZ to Authoring Tool:
  - Feltrim et al. (Sao Paulo University): ported AZ to Portugese
  - AZ student's introductions of CS theses and critique structure
- Application of AZ to different . . .
  - text type (CS journal articles);
  - language (Portugese)
  - domains (chemistry, biology; projects SciBorg, FlySlip); use of AZ to pinpoint location for IE
- New features for Citation Function Classification (Siddharthan & Teufel, 2007, HLT/NAACL)
- Automatic aquisition of cues for AZ (Abdalla & Teufel, 2006, ACL)

Abdalla and Teufel (2006): Meta-Discourse Detection

Correctly found:

> *What we aim in this paper is to propose a paradigm that enables partial / local generation through decompositions and reorganizations of tentative local structures.*

Correctly rejected:

> *Perhaps the method proposed by Pereira et al. (1993) is the most relevant in our context.*

References

Abdalla, R., S. Teufel (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. In: *Proc. of ACL-06*, Sydney, Australia.

Feltrim, V., S. Teufel, G. Gracas Nunes and S. Alusio (2005). Argumentative Zoning applied to Critiquing Novices' Scientific Abstracts In *Computing Attitude and Affect in Text: Theory and Applications* Shanahan, J.; Qu, Y.; Wiebe, J. (Eds.) 2005, Spinger.

Siddharthan, A. and S. Teufel (2007). Whose idea was that? Determining Scientific Attribution. To Appear in: *Proc. of NAACL/HLT-07, Rochester, New York.*

Swales, J. (1990). In: *Genre Analysis: English in Academic and Research Settings.* Cambridge University Press, UK.

Teufel, S., A. Siddharthan and D. Tidhar (2006a). Automatic classification of citation function. In: *Proc. of EMNLP-06*, Sydney, Australia.

Teufel, S., A. Siddharthan and D. Tidhar, (2006b). An annotation scheme for citation function. In: *Proc. of SigDial-06*, Sydney, Australia.

Teufel, S., M. Moens (2002). Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

Teufel, S. (2001). Task-based evaluation of summary quality: Describing relationships between scientific papers. In: *Proc. Automatic Text Summarization Workshop at NAACL-01*, Pittsburgh, PA.

Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*, PhD. Thesis, University of Edinburgh, UK.

Teufel, S., J. Carletta, M. Moens (1999). An annotation scheme for discourse-level argumentation in research articles. In: *Proc. of EACL-99*, Bergen, Norway.

Examples of System classifications (CitRAZ)

Context	Human	Machine	Comment
We have compared four complete and three partial data representation formats for the baseNP recognition task presented in **Ramshaw and Marcus (1995)**.	PUse	PUse	weak cues
In the version of the algorithm that we have used, IB1-IG, the distances between feature representations are computed as the weighted sum of distances between individual features **(Bosch 1998)**.	Neut	PUse	detail in used package, not really used.
We have used the baseNP data presented in **Ramshaw and Marcus (1995)**.	PUse	PUse	Straightforw.
We will follow **Argamon et al. (1998)** and use a combination of the precision and recall rates: F=(2*precision*recall)/(precision+recall).	PSim	PUse	F-measure not attributable to cit.
This algorithm standardly uses the single training item closest to the test i.e. However **Daelemans et al. (1999)** report that for baseNP recognition better results can be obtained by making the algorithm consider the classification values of the three closest training items.	Neut	PUse	Machine mislead by cue.
They are better than the results for section 15 because more training data was used in these experiments. Again the best result was obtained with IOB1 (F=92.37) which is an improvement of the best reported F-rate for this data set (**(Ramshaw and Marcus 1995)** (F=92.03).	CoCo-	PUse	Machine misled by cue.

Human AZ annotation

- Human annotation (Teufel, Carletta and Moens 1999, EACL)
  - 3 trained annotators, 25 articles
  - high agreement: K = .71 (inter); K = .82, .81, .76 (intra)
  - Skewed distribution (70% OWN, 2% AIM)
- What did the annotators have problems with?
  - OWN v. OTHER in case of “close” previous work
  - OTHER v. BACKGROUND is a question of degree
  - AIM v. CONTRAST in long sentences
  - CONTRAST v. OTHER in case of non-overt criticism

Example of an “AZ-extract”

AIM

22 We now give a similarity-based method for estimating the probabilities of cooccurrences unseen in training.

151 Our method combines similarity-based estimates with Katz’s back-off scheme, which is widely used for language modeling in speech recognition.

CONTRAST

20 Their model, however, is not probabilistic, that is, it does not provide a probability estimate for unobserved cooccurrences.

28 We applied our method to estimate unseen bigram probabilities for Wall Street Journal text and compared it to the standard back-off model.

115 We will outline here the main parallels and differences between our method and cooccurrence smoothing.

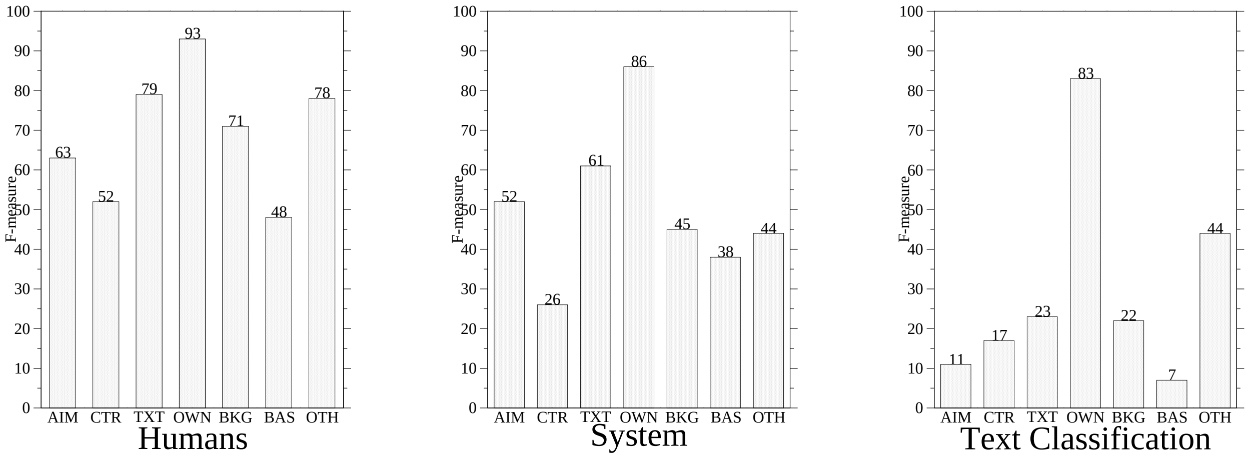
BASIS

23 Similarity-based estimation was first used for language modeling in the cooccurrence smoothing method of Essen and Steinbiss (1992), derived from work on acoustic model smoothing by Sugawara et al. (1985).

87 The baseline back-off model follows closely the Katz design, except that for compactness all frequency one bigrams are ignored.

122 Notice that this formula has the same form as our similarity model CREF, except that it uses confusion probabilities where we use normalized weights.

Intrinsic evaluation: comparison to human annotation



	Humans	System	Baselines:			
			T. Class.	Rand.	R. (Distr.)	Most Freq.
Accuracy	.87	.73	.72	.14	.48	.67
Macro-F	.69	.50	.30	.09	.14	.11
Kappa	.71	.45	.30	–.10	0	–.13

- System beats all baselines (Teufel & Moens, 2002)
- Human–system agreement not high (K=.45)

Intrinsic evaluation: confusion matrix

		H 2							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
	AIM	35	2	1	19	3		2	62
	CTR		86		31	16		23	156
	TXT			31	7			1	39
H 1	OWN	10	62	5	2298	25	3	84	2487
	BKG		5		13	115		20	153
	BAS	2			18	1	18	14	53
	OTH	1	18	2	55	10	1	412	499
	Total	48	173	39	2441	170	22	556	3449

Human–human

		M							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
	AIM	127	6	13	23	19	5	10	203
	CTR	21	112	4	204	87	18	126	572
	TXT	14	1	145	46	6	2	6	220
H	OWN	100	108	84	7231	222	71	424	8240
	BKG	14	31	1	222	370	5	101	744
	BAS	17	7	7	60	8	97	39	235
	OTH	6	70	10	828	215	72	773	1974
	Total	299	335	264	8614	927	270	1479	12188

Human–machine

Comparison of example AZ-extract to Human Gold Standard

AIM			
	22	We now give a similarity-based method for estimating the probabilities of cooccurrences unseen in training.	√
	151	Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition.	BASIS
CONTRAST			
	20	Their model, however, is not probabilistic, that is, it does not provide a probability estimate for unobserved cooccurrences.	√
	28	We applied our method to estimate unseen bigram probabilities for Wall Street Journal text and compared it to the standard back-off model.	OWN
	115	We will outline here the main parallels and differences between our method and cooccurrence smoothing.	√
BASIS			
	23	Similarity-based estimation was first used for language modeling in the cooccurrence smoothing method of Essen and Steinbiss (1992), derived from work on acoustic model smoothing by Sugawara et al. (1985).	OTHER
	87	The baseline back-off model follows closely the Katz design, except that for compactness all frequency one bigrams are ignored.	√
	122	Notice that this formula has the same form as our similarity model CREF, except that it uses confusion probabilities where we use normalized weights.	CONTR.

Extrinsic evaluation

- Teufel (2001, WS-sum)
- Task: users are asked to pick from the reference list those papers which have a **contrast** or a **research continuity** to the current paper (and state what the contrast/continuity is)
- 24 subjects, 6 experimental groups, 6 randomly chosen articles
- Compare 6 conditions:

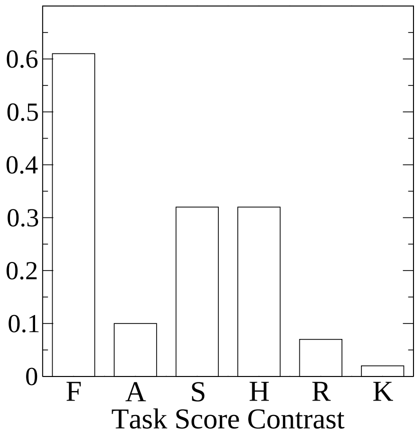
A	Author abstract	Controlled for length
K	List of keywords (TF*IDF)	
R	Random sentences	
S	AZ-extract (system output)	
H	AZ-extract (human generated)	
F	Full article	

- Gold standards are taken from group with full papers

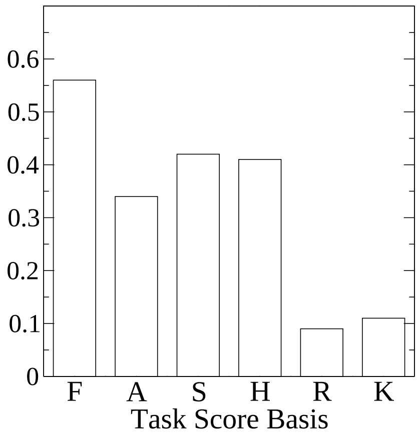
- Give scores for right answers, weight by number of judges who agree

Extrinsic evaluation results: task scores

CONTRAST



BASIS



Indistinguishable ($p < .01$) (Wilcoxon matched-pairs signed-rank test):

S–H
A–K
K–R
A–R (Contrast only)

- AZ-extracts beat all other short conditions
- System output statistically indistinguishable from human output

Conclusion

- Document structure recognition to support document management tasks
- Argumentative Zoning (robust rhetorical analysis) and citation function classification
  - Can both be done reliably by humans and okay by machines
- Applications of AZ/CFC:
  - AZ-extracts shown to be useful in information gathering task
  - Citation maps
  - Authoring tool
- Methodology from applied fields helps to substantiate AZ's theoretical claims

CFC: Results, Collapsing Categories

	Weakness	Positive	Contrast	Neutral
P	.80	.75	.77	.81
R	.49	.65	.52	.90
F	.61	.70	.62	.86

Percentage Accuracy	0.79
Kappa (n=12; N=2829; k=2)	0.59
Macro-F	0.68

	Weakness	Positive	Neutral
P	.77	.75	.85
R	.42	.65	.92
F	.54	.70	.89

Percentage Accuracy	0.83
Kappa (n=12; N=2829; k=2)	0.58
Macro-F	0.71

Naïve Bayesian Classifier

Naïve Bayesian model (from Kupiec, Pedersen, Chen, 1995).

$$P(C|F_0, \ldots, F_{n-1}) \approx \quad P(C)\ \frac{\Pi_{j=0}^{n-1} P(F_j|C)}{\Pi_{j=0}^{n-1} P(F_j)}$$

$P(C|F_0, \ldots, F_{n-1})$: Probability that a sentence has target category C, given its feature values F_0, . . . , F_{n-1};

$P(C)$: (Overall) probability of category C;

$P(F_j|C)$: Probability of feature-value pair F_j, given that the sentence is of target category C;

$P(F_j)$: Probability of feature value F_j;

Example of an answer, scored

AIM
Extending co-occurrence probabilities of unseen events using similarity measures and a corpus

CONTRAST

?	not probabilistic	1 (half)
Cooccurence smoothing (Essen, Steinbiss, 1992)	differences	3
Katz 87, standard back-off model	differences	1
		5/8

BASIS

Katz 87 back-off model	further development	3
Essen and Steinbiss 92	idea and formula	1
		4/10