Corpora for conceptualisation and zoning of scientific papers

Maria Liakata[1], Simone Teufel[2], Advaith Siddharthan[3], Colin Batchelor[4]

[1]Department of Computer Science, Aberystwyth University, mal@aber.ac.uk
[2]Computer Laboratory, University of Cambridge, Simone.Teufel@cl.cam.ac.uk
[3]Department of Computing Science, University of Aberdeen, advaith@abdn.ac.uk
[4]Informatics R&D, Royal Society of Chemistry, batchelorC@rsc.org

Abstract

We present two complementary annotation schemes for sentence based annotation of full scientific papers, CoreSC and AZ-II, which have been applied to primary research articles in chemistry. The AZ scheme is based on the rhetorical structure of a scientific paper and follows the knowledge claims made by the authors. It has been shown to be reliably annotated by independent human coders and has proven useful for various information access tasks. AZ-II is its extended version, which has been successfully applied to chemistry. The CoreSC scheme takes a different view of scientific papers, treating them as the humanly readable representations of scientific investigations. It therefore seeks to retrieve the structure of the investigation from the paper as generic high-level Core Scientific Concepts (CoreSC). CoreSCs have been annotated by 16 chemistry experts over a total of 265 full papers in physical chemistry and biochemistry. We describe the differences and similarities between the two schemes in detail and present the two corpora produced using each scheme. There are 36 shared papers in the corpora, which allows us to quantitatively compare aspects of the annotation schemes. We show the correlation between the two schemes, their strengths and weaknesses and discuss the benefits of combining a rhetorical based analysis of the papers with a content-based one.

1. Introduction

Annotation schemes and corpora for scientific texts, especially in the biomedical domain, are becoming increasingly important in enabling the automatic processing of information. Such schemes look at annotating mostly abstracts of papers and less often full papers, with the majority focussing on annotation at the token level for keywords, (Korhonen et al., 2009; Thompson et al., 2009). However, many consider more complex linguistic phenomena such as negation, hedges, dependencies and semantic relations at either the token or sentence level (Vincze et al., 2008; Medlock and Briscoe, 2007; McIntosh and Curran, 2009) and at the sentence level for discourse-based categories (Hirohata et al., 2008; Teufel et al., 2009).

In the following we present and compare two complementary sentence-based annotation schemes, CoreSC and AZ-II, which we have used to annotate full scientific papers in chemistry.

2. The CoreSC scheme

2.1. Core Scientific Concepts

The CoreSC annotation scheme adopts the view that a scientific paper is the human-readable representation of a scientific investigation and therefore seeks to mark the components of a scientific investigation as expressed in the text. CoreSC is ontology-motivated and originates from the CISP meta-data (Soldatova and Liakata, 2007), a subset of classes from EXPO (Soldatova and King, 2006), an ontology for the description of scientific investigations. CISP consists of the concepts: `Motivation`, `Goal`, `Object`, `Method`, `Experiment`, `Observation`, `Result` and `Conclusion`, which were validated using an on-line survey as constituting the indispensable set of concepts necessary for the description of a scientific investigation. CoreSC implements these as well as `Hypothesis`, `Model` and `Background`, as a sentence-based annotation scheme for 3-layered annotation. The first layer pertains to the previously mentioned 11 categories, the second layer is for the annotation of properties of the concepts (e.g. "New", "Old") and the third layer caters for identifiers (conceptID), which link together instances of the same concept, e.g. all the sentences pertaining to the same method will be linked together with the same conceptID (e.g. "Met1").

If we combine the layers of annotation so as to give flat labels, we cater for the categories in table 1.

The CoreSC scheme was accompanied by a set of 45 page guidelines which contain a decision tree, detailed description of the semantics of the categories, 6 rules for pairwise distinction and examples from chemistry papers. These guidelines are available from http://ie-repository.jisc.ac.uk/88/.

2.2. The CoreSC corpus

We used the CoreSC annotation scheme and the semantic annotation tool SAPIENT (Liakata et al., 2009) to construct a corpus of 265 annotated papers (Liakata and Soldatova, 2009) from physical chemistry and biochemistry. The CoreSC corpus was developed in two different phases. During phase I, fifteen Chemistry experts were split into five groups of three, each of which annotated eight different papers; A 16th expert annotated across groups as a consistency check. This resulted in a total of 41 papers being annotated, all of which received multiple annotations. We ranked annotators according to median success in terms of inter-annotator agreement (as measured by Cohen's(Cohen, 1960) kappa) both within their groups and for a paper common across groups. In phase II, the 9 best annotators of phase I each annotated 25 papers, amounting to a total of 225 papers.

Table 1: The CoreSC Annotation scheme

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	the data/phenomena recorded in an investigation
Result	factual statements about the outputs of an investigation
Conclusion	statements inferred from observations & results relating to research hypothesis

Table 2: AZ-II Annotation Scheme.

Category	Description	Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper	OWN_CONC	Findings, conclusions (non-measurable) of own work
NOV_ADV	Novelty or advantage of own approach	CODI	Comparison, contrast, difference to other solution (neutral)
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)	GAP_WEAK	Lack of solution in field, problem with other solutions
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description	ANTISUPP	Clash with somebody else's results or theory; superiority of own work
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.	SUPPORT	Other work supports current work or is supported by current work
OWN_MTHD	New Knowledge claim, own work: methods	USE	Other work is used in own work
OWN_FAIL	A solution/method/experiment in the paper that did not work	FUT	Statements/suggestions about future work (own or general)
OWN_RES	Measurable/objective outcome of own work		

Evaluation statistics from phase I are presented in section 5.1. Work in progress involves evaluation of annotation in phase II. The corpus can be downloaded from:
http://www.aber.ac.uk/en/ns/research/cb/projects/art/art-corpus/

3. The AZ-II annotation scheme

3.1. The AZ-II categories

In contrast to CoreSC, the AZ-II annotation scheme (Table 2) models rhetorical and argumentational aspects of scientific writing, and in particular concentrates on rhetorical statements and on connections between the current paper and cited papers (there are two roughly negative or adversarial categories (GAP_WEAK and ANTISUPP), a neutral one that marks contrasts (CODI) and several positive ones (USE and SUPPORT).

The main ordering principle of AZ-II is based on who a given knowledge claim belongs to: the authors of the paper, some other cited work (OTHR), or nobody in general (CO_GRO). The majority of the paper is reserved for the new knowledge claim that the authors are trying to defend in the paper. These parts of the paper are shared between the categories OWN_MTHD (methods), OWN_RES (results), OWN_CONC (conclusion), and local failure (OWN_FAIL).

3.2. The AZ-II corpus

The AZ-II annotated corpus consists of 61 articles from the Royal Society of Chemistry. 30 of the papers are annotated by three annotators; the remaining ones by one or two annotators. Reliability was measured on the 30 papers. More information can be found in (Teufel et al., 2009).

4. CoreSC and AZ-II comparison

The two schemes are complementary in that they take different views on what a scientific paper represents. AZ as-

Table 3: Summary Table of category distributions and annotation performance for CoreSC categories in phase I

Category	Freq.	(Cohen κ)	(Byrt κ)	Category	Freq.	(Cohen κ)	(Byrt κ)
Conclusion	10.56%	**0.89**	**0.79**	Experiment	16.8%	**0.65**	**0.58**
Background	16.6%	**0.87**	**0.75**	Goal	1.82%	**0.6**	**0.58**
Observation	13.68%	**0.79**	**0.67**	Hypothesis	2.39%	**0.46**	**0.44**
Object	3.48%	**0.81**	**0.77**	Motivation	2.25%	**0.46**	**0.44**
Result	18.51%	**0.78**	**0.6**	Model	5.34%	**0.43**	**0.39**
Method	9.82%	**0.74**	**0.6**				

sumes that a paper is the attempt of claiming ownership for a new piece of knowledge and aims to recover the rhetorical structure and the relevant stages in this argument.

CoreSC on the other hand treats scientific papers as the humanly readable representations of scientific investigations. It therefore seeks to retrieve the structure of the investigation from the paper in the form of generic high-level Core Scientific Concepts. Thus, they have different focus with CoreSC containing more categories pertaining to the content of the paper whereas AZ categories elaborate on the path to various knowledge claims.
The two schemes also differ in that CoreSC so far has used expert knowledge for annotation, whereas AZ-II has been annotated by expert-trained non-experts in a procedure specified in (Teufel et al., 2009). The schemes have common ground, in the sense that they are both sentence based and target scientific papers. They even share some category names in common, such as "Method", "Result" and "Conclusion", even though these are defined differently in the two schemes and differ in granularity.

More specifically, Background in CoreSC covers generally accepted neutral background knowledge but also existing knowledge claims, represented in AZ-II through the OTHR, PREV_OWN and CO_GRO categories. The AIM category in AZ-II is a statement of research goal; In CoreSC this is split into three categories: Goal (the target state of the investigation), Hypothesis (a statement not yet confirmed) and Object (a statement pertaining to a particular entity-product of the investigation). Object, though, can also refer to any statement assigning novelty or advantage properties to a principle entity of the investigation. OWN_MTHD and Method both refer to methods used. However, CoreSC allows the distinction into experimental method (Experiment), other types of methods used in the current work Method-New and methods used in other work mentioned in the paper (Method-Old). OWN_RES corresponds to the CoreSC category Observation, which represents the data/phenomena recorded within an investigation. By contrast, the CoreSC category Result pertains to the factual statements derived from Observation. Conclusion in CoreSC involves statements inferred from observations and results, relating to the Hypothesis. AZ-II contains a category called NOV_ADV, which stands for the novelty or advantage of the approach mentioned in the paper. In CoreSC, one can annotate the novelty and advantage of Method and Object. The rest of the categories are completely distinct for the two schemes. In CoreSC, Hypothesis, Motivation, Object and Model complete the underlying investigation structure whereas in AZ-II CODI, GAP_WEAK, SUPPORT, ANTISUPP, USE and FUT follow the connection with other work and OWN_FAIL expresses local failure.

5. Annotation results

5.1. CoreSC annotation results

The inter-annotator agreement presented here for CoreSC is based on phase I of the corpus development (41 papers). Work in progress involves evalutation of papers in phase II (225 papers). The inter-annotator agreement for the 9 best performing annotators was κ= 0.57 for the paper common across all annotators (N=255,n=11,k=9)[1]. For the rest of the papers, the inter-annotator agreement was κ=0.5 (N=5022,n=11,k=9). The score we report is Cohen's κ (Cohen, 1960), but κ calculated according to Siegel and Castellan's (1988) formula were very similar. As the quality of annotators was determined post-hoc, it is independent of group assignment. Hence, groups often consist of more and less reliable annotators. Thus, the κ score is based on an annotator's agreement within their group, which often consisted of reliable and less reliable annotators.
The frequency and distinguishability of categories is given in Table 3. Result and Experiment are the most frequent categories at roughly 17–18.5%, the 5 least frequent categories taken together (Goal, Hypothesis, Motivation, Model, Object) make up 15.28% of the corpus. Distinguishability was measured according to Krippendorff's (1980) diagnostic, which collapses all categories but the one in focus into one category and then measures reproducibility. If it goes up significantly, this category is better distinguished than the overall distinction of categories. We report Cohen's κ and Byrt's (Byrt et al., 1993) κ. Conclusion, Background,Observation and Object are easier to recognise, whereas Hypothesis, Motivation and Model are harder to recognise than the average taken at κ=0.55.

5.2. AZ-II annotation results

The inter-annotator agreement for the AZ-II corpus was κ = 0.71 (N=3745,n=15,k=3), here reported in terms of the (Fleiss, 1971) κ. The frequency and distinguishability of categories is given in Table 5.1.. OWN_MTHD and

[1]N stands for the number of sentences, n for the number of categories and k for the number of annotators

Table 4: Frequency and Annotation Performance of AZ-II Categories.

Category	Freq.	Perf. (κ)	Category	Freq.	Perf. (κ)
OWN_MTHD	25.4%	**0.76 ± 0.03**	SUPPORT	1.5%	0.67 ± 0.15
OWN_RES	24.0%	0.73 ± 0.03	GAP_WEAK	1.1%	0.63 ± 0.17
OWN_CONC	15.1%	**0.63 ± 0.04**	FUT	1.0%	0.72 ± 0.18
OTHR	8.3%	0.65 ± 0.06	NOV_ADV	1.0%	0.64 ± 0.18
USE	7.9%	**0.82 ± 0.06**	CODI	0.8%	**0.35 ± 0.19**
CO_GRO	6.7%	0.69 ± 0.07	OWN_FAIL	0.8%	0.52 ± 0.20
PREV_OWN	3.4%	**0.60 ± 0.10**	ANTISUPP	0.5%	**0.36 ± 0.26**
AIM	2.3%	0.80 ± 0.12			

Table 5: Contingency table for CoreSC and AZ-II

CoreSC	AIM	ANTISUPP	CODI	CO_GRO	FUT	GAP-WEAK	NOV-ADV	OTHR	OWN-CONC	OWN-FAIL	OWN-MTHD	OWN-RES	PREV_OWN	SUPP	USE	0	**Total**
Background	7	3	13	259	17	22	10	200	55	3	93	23	77	6	11	74	873
Conclusion	22	5	15	6	28	2	26	5	278	8	16	32	2	15	0	25	485
Experiment	1	0	1	0	0	0	0	10	5	3	608	39	10	1	184	21	883
Goal	23	0	0	1	1	0	5	0	4	0	29	3	1	0	2	1	70
Hypothesis	2	0	0	5	4	0	1	3	59	0	21	0	1	0	0	4	100
Method-New	5	0	3	0	1	0	1	1	8	0	124	5	2	0	20	13	183
Method-New-Adv	0	1	1	1	0	0	6	0	4	1	19	0	0	0	0	0	33
Method-New-Dis	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	1	5
Method-Old	0	0	0	19	0	2	0	21	1	0	30	5	15	0	30	6	129
Method-Old-Adv	0	0	0	8	0	0	4	1	0	0	0	0	1	0	0	0	14
Method-Old-Dis	0	0	0	4	0	5	0	0	1	1	0	0	1	0	0	0	12
Model	0	0	0	11	1	0	0	12	43	2	86	23	3	0	15	14	210
Motivation	0	0	0	38	0	13	2	1	2	0	4	0	4	0	0	2	66
Object-New	31	1	0	2	0	0	1	0	2	0	50	3	1	1	3	18	113
Object-New-Adv	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	3
Observation	3	1	2	2	0	0	0	2	27	6	45	414	3	2	4	137	648
Result	17	6	13	2	3	1	4	5	203	7	68	407	5	18	4	68	831
Total	112	17	48	359	55	45	60	261	694	31	1196	954	126	43	273	384	4658

OWN_RES are the most frequent categories at roughly 24–25%, whereas the 7 least frequent categories taken together (SUPPORT, GAP_WEAK, FUT, NOV_ADV, CODI, OWN_FAILand ANTISUPP) only make up 6.7% of the corpus. Distinguishability was measured according to Krippendorff's diagnostic. Significance was measured using the (Fleiss et al., 1969) formula; boldfaced numbers indicate significantly better or worse performance. OWN_MTHD and USE are significantly easier to recognise, whereas OWN_CONC, PREV_OWN, CODI and ANTISUPP are significantly harder to recognise than the average AZ-II category.

6. The joint CoreSC-AZ set

6.1. Measured association between the two schemes

36 papers have been annotated both with AZ-II and CoreSC. This allowed us to calculate the correlation between the two schemes as reflected within the papers, map categories between schemes and assess the level of their complementarity. For instance, we were able to find the relation between OWN_RES and `Observations`, OWN_CONC vs `Results`, `Conclusions` and how the knowledge claim AZ categories are distributed among the rest of the content-based CoreSC categories.

Table 5 gives the contingency matrix between the two schemes, where the CoreSC categories constitute the rows and the AZ-II categories constitute the columns.

As a first step we wanted to assess whether there is a statistically significant correlation between the two schemes, that is, whether there is some sort of association between the rows and columns of the contingency table. We calculated the chi-squared Pearson statistic and the chi-squared likelihood ratio both of which showed a definite association between CoreSC and AZ-II categories. This result was further backed by the values of the contingency coefficient and Cramer's V (Table 6)[2].

However, these measures cannot give an indication of the degree of association between the two schemes or whether the association is symmetric, i.e. whether it goes both directions and to what extent. To obtain a measure for the differential association between CoreSC and AZ-II we calculated the Goodman-Kruskal lambda L statistic (Siegel and Castellan, 1988), which gives us the reduction in error for predicting the categories of one annotation scheme, if we know the categories assigned according to the other scheme. When using AZ-II (columns) as the independent variable, we obtained a Goodman-Kruskal lambda of 0.38, which means that knowing the AZ-II categories assigned would help us reduce the error in predicting the CoreSC categories by 38%. When using CoreSCs as the independent variable, the reduction in prediction error of AZ-II categories given CoreSC categories was calculated as 35%, according to the Goodman Kruskal lambda. To test the sig-

[2]These are association measures for r x c tables. We used the implementation in the vcd package of R (http://www.r-project.org/).

Table 6: Association measures between CoreSC and AZ-II

	X^2	df	$P(> X^2)$
Likelihood Ratio	6360.8	288	0
Pearson	8372.0	288	0
Contingency Coeff		0.802	
Cramer's V		0.335	

nificance of the lambda statistic we tested the hypotheses that the reduction in prediction error is $H_{01} : l = .33$ and $H_{02} : l = .35$ (See table 7). Since the number of sentences is relatively large (4658) we can assume that L follows a normal distribution. Calculating z showed that we could assume at a significance level of $a = 0.01$ that we could reject the null hypothesis and assume that $l >= 0.35$ and $l >= 0.33$ when AZ-II and CoreSC were chosen respectively as the independent category.

We then interpreted the contingency table in order to show the correlation between the two schemes in terms of the actual categories. Each category in the CoreSC scheme is expressed as a percentage of the AZ-II categories and vice versa. This is reflected in tables 8 and 9 respectively.

6.2. Discussion of correlated categories

Looking at Table 8 we can see that >61% of the CoreSC `Background` category corresponds to AZ-II categories which pertain to previous work: CO_GRO (29.5%), OTHR (22.9%), PREV_OWN(8.8%) . It seems that 10.65% of it is OWN_MTHD, and another 8.4% has not been annotated at all by AZ-II. There are other AZ-II categories that contribute to the total for `Background` at smaller percentages. This is in accordance with our expectations as from its definition `Background` is a rather broad category (See section 4.). The majority of the CoreSC category `Conclusion` corresponds to the AZ-II categories OWN_CONC (57.31%) with OWN_RES(6.59%), FUT (5.77%), NOV_ADV (5.36%) and 5.15% corresponding to sentences not annotated by AZ-II. The majority overlap is encouraging, although the percentages assigned to OWN_RES and the unannotated 5.15% suggest some disagreement between the two schemes as to what counts as a result and what counts as a conclusion. The other categories such as FUT and NOV_ADV both pertain to final outcomes of the investigation, which according to the definition of the CoreSC scheme are expected to count as `Conclusion`.

The CoreSC category `Experiment` seems to consist primarily of OWN_MTHD (68.85%) and USE (20.8%). This is to be expected as `Experiment` encodes experimental methods in particular. The `Goal` category consists primarily in OWN_MTHD(41.4%) and AIM(32.86%). The overlap with AIM is no surprise, as according to section 4., by definition we expect AIM to be split into `Goal`, `Object` and `Hypothesis`. However, the percentage covered by OWN_MTHD perhaps suggests the overgeneral nature of the AZ-II category OWN_MTHD. The CoreSC `Hypothesis` seems to be 59% OWN_CONC and OWN_MTHD 21%. While some overlap with OWN_CONC could be expected, this result suggests the need for more fine-grained characterisation of OWN_CONC and OWN_MTHD.

The CoreSC `Method` and its properties encouragingly correspond mostly to the OWN_MTHD AZ-II category. More specifically, `Method-New` is 67.75% OWN_MTHD 10.9% USE and 7.1% unassigned. `Method-New-Advantage` is 57.57% OWN_MTHD and 18.1% NOV_ADV. The latter is positive, as both schemes seem to agree on the novelty-advantage of the method. `Method-New-Disadvantage` is partitioned between OWN_MTHD (40%), OWN_CONC (40%) and unassigned (20%). `Method-Old` is a combination of OWN_MTHD (23.25%), USE (23.25%), OTHR (16.27%), CO_GRO (14.7%), PREV_OWN (11.6%). `Method-Old-Advantage` is 57.14% CO_GRO and 28.43% NOV_ADV whereas interestingly `Method-Old-Disadvantage` is 40.95% GAP_WEAK and 33.3% CO_GRO.

The CoreSC `Model`, which essentially corresponds to theoretical methods and assumptions, is expressed as OWN_MTHD(40.95%), OWN_CONC(20.47%), OWN_RES(10.95%), USE (7.14%) and 6.66% unassigned.

`Motivation` is primarily CO_GRO (57.57%) and GAP_WEAK (19.69%). `Object-New` is split between OWN_MTHD (44.2%), AIM (27.43%) and 15.9% unassigned. These percentages are consistent with our expectations as often methods are objects of the investigation. `Object-New-Advantage` is equally split between AIM, CO_GRO and OWN_MTHD but there are not enough instances to make this a meaningful observation.

The CoreSC `Observation` primarily maps to AZ-II OWN_RES (63.88%) but also 21.1% of it remains unassigned by AZ-II. `Result` consists in OWN_RES (48.97%), OWN_CONC (24.4%), OWN_MTHD (8.2%) whereas another 8.2% remains unassigned. So it seems that the CoreSC `Observation` is closer to AZ-II OWN_RES than `Result` actually is. In CoreSC we make a distinction between raw data observations `Observation`, intermediate results `Result` and final `Conclusions` whereas AZ-II only distinguishes measurable/objective outcomes (OWN_RES) and non-measurable outcomes (OWN_CONC).

Table 9 expresses the AZ-II categories in terms of the CoreSC annotation scheme. AIM is expressed as 27.67% `Object-New`, 20.53% `Goal`, 19.64% `Conclusion` and 15.18% `Result`. This wide range of categories corresponding to AIM is somewhat to be expected as it is by definition[3] a broad category in terms of CoreSCs. OWN_CONC is mostly `Conclusion`, `Result` and `Hypothesis` (40.06%, 29.25% and 8.5% respectively and a range of other categories at smaller percentages) whereas OWN_RES is 43.4% `Observation` and `Result` 42.66%. This shows difference in granularity between the two pairs of concepts, OWN_CONC vs `Conclusion` and OWN_RES vs `Result`. The

[3] AIM is defined as "Statement of specific research goal or hypothesis of current paper". Wel also saw that `Hypothesis` corresponds primarily to OWN_CONC which in turn is mostly `Conclusion` and `Result`

Table 7: Differential association between CoreSC and AZ-II

	L	$P(Ho : l = .33)$	$P(Ho : l = .35)$
Goodman-Kruskal lambda (cols/AZ-II)	0.3793377	<0.00003	0.00023
Goodman-Kruskal lambda (rows/CoreSC)	0.3523975	0.0099	0.4013

Table 8: CoreSC categories expressed in terms of AZ-II

CoreSC	AZ Category distribution in (%)
`Background`	CO_GRO 29.6% OTHR 22.9% OWN_MTHD 10.65% PREV_OWN 8.8% 0 8.4% OWN_CONC 6.3% OWN_RES 2.6% GAP_WEAK 2.5% FUT 1.9% USE 1.26% NOV_ADV 1.13% AIM 0.8% SUPPORT 0.68% (OWN_FAIL , ANTISUPP) 0.34%
`Conclusion`	OWN_CONC 57.31% OWN_RES 6.59% FUT 5.77% NOV_ADV 5.36% 0 5.15% AIM 4.53% OWN_MTHD 3.3% (CODI , SUPPORT) 3.1% OWN_FAIL 1.64% CO_GRO 1.23% (ANTISUPP , OTHR) 1% (GAP_WEAK , PREV_OWN) 0.41%
`Experiment`	OWN_MTHD 68.85% USE 20.8% OWN_RES 4.41% 0 2.37% (OTHR , PREV_OWN) 1.1% OWN_CONC 0.56% OWN_FAIL 0.33% (AIM , SUPPORT , CODI) 0.11%
`Goal`	OWN_MTHD 41.4% AIM 32.86% NOV_ADV 7.1% OWN_CONC 5.71% OWN_RES 4.28% USE 2.85% (CO_GRO , FUT , PREV_OWN ,0) 1.4%
`Hypothesis`	OWN_CONC 59% OWN_MTHD 21% CO_GRO 5% (FUT ,0) 4% OTHR 3% AIM 2% (NOV_ADV , PREV_OWN) 1%
`Method-New`	OWN_MTHD 67.75% USE 10.9% 0 7.1% OWN_CONC 4.37% (AIM OWN_RES) 2.73% CODI 1.64% PREV_OWN 1% (FUT , NOV_ADV , OTHR) 0.5%
`Method-New-Advantage`	OWN_MTHD 57.57% NOV_ADV 18.1% OWN_CONC 12.1% (ANTISUPP , CODI , CO_GRO , OWN_FAIL) 3%
`Method-New-Disadvantage`	OWN_MTHD 40% OWN_CONC 40% 0 20%
`Method-Old`	(OWN_MTHD , USE) 23.25% OTHR 16.27% CO_GRO 14.7% PREV_OWN 11.6% 0 4.6% OWN_RES 3.87% GAP_WEAK 1.5% OWN_CONC 0.77%
`Method-Old-Advantage`	CO_GRO 57.14% NOV_ADV 28.43% (OTHR , PREV_OWN) 7.1%
`Method-Old-Disadvantage`	GAP_WEAK 41.66% CO_GRO 33.3% (OWN_CONC , OWN_FAIL , PREV_OWN) 8.33%
`Model`	OWN_MTHD 40.95% OWN_CONC 20.47% OWN_RES 10.95% USE 7.14% 0 6.66% OTHR 5.71% CO_GRO 5.23% PREV_OWN 1.42% OWN_FAIL 0.95% FUT 0.47%
`Motivation`	CO_GRO 57.57% GAP_WEAK 19.69% (OWN_MTHD , PREV_OWN) 6% (SUPPORT , OWN_CONC ,0) 3% OTHR 1.5 %
`Object-New`	OWN_MTHD 44.2% AIM 27.43% 0 15.9% (OWN_RES , USE) 2.65% (CO_GRO , OWN_CONC) 1.76% (ANTISUPP ,SUPPORT , NOV_ADV , PREV_OWN) 0.88%
`Object-New-Advantage`	(AIM , CO_GRO , OWN_MTHD) 33.33%
`Observation`	OWN_RES 63.88% 0 21.1% OWN_MTHD 6.94% OWN_CONC 4.16% OWN_FAIL 0.93% USE 0.62% (AIM , PREV_OWN) 0.46% (CODI , CO_GRO , OTHR , SUPPORT) 0.3% ANTISUPP 0.15%
`Result`	OWN_RES 48.97% OWN_CONC 24.4% (OWN_MTHD ,0) 8.2% SUPPORT 2.16% AIM 2% CODI 1.56% OWN_FAIL 0.84% ANTISUPP 0.72% (OTHR , PREV_OWN) 0.6% (NOV_ADV , USE) 0.48% FUT 0.36% CO_GRO 0.24% GAP_WEAK 0.12%

same holds for OWN_MTHD and the CoreSC `Method`, since OWN_MTHD is expressed as 50.84% `Experiment`, 14.64% `Method`(incl. properties), 7.78% `Background`, 7.19% `Model`, 5.68% `Result`, 4.18% `Object-New`. NOV_ADV(*novelty or advantage of own approach*) is 49.06% `Conclusion`, 15.09% `Background`, 11.32% `Method-New-Advantage`, 9.43% `Goal`. This is similar in principle to `Method-New-Advantage` (hence the overlap), but is broader as it is not necessarily confined to methods. GAP_WEAK(*lack of solution in field*) is 48.88% `Background`, 28.88% `Motivation`, 11.11% `Method-Old-Disadvantage`, 4.44% `Method-Old`, 4.44% `Conclusion`). The overlap with `Motivation` and `Method-Old-Disadvantage` is encouraging with respect to the semantics of the two schemes.

The rest of the AZ-II categories permeate across difference CoreSC concepts, which is what we expect since by design they follow the progress of knowledge claims. For example, PREV_OWN (*knowledge claim held by the author's in previous paper*) is 61.1% `Background` and 11.9% `Method-Old`. CO_GRO(*No knowledge claim*) is 72.06% `Background`, 10.61% `Motivation`, 8.64% `Method-Old` and 3% `Model`. OTHR(*Significant knowledge claim made by other researchers*) is 76.63% `Background`, 8.05% `Method-Old`, 4.6% `Model`, 3.83% `Experiment`. USE(*other work used in own work*) is 67.4% `Experiment`, 10.99% `Method-Old`, 7.33% `Method-New`, 5.49% `Model`. CODI(*neutral comparison to other work*) is 31.25% `Conclusion`, 27.1% `Result`, 27.1% `Background`, 8.35% `Method-New`, 4.16% `Observation`. SUPPORT(*other work supports current work*) is 43.9% `Result`, 36.59%

Table 9: AZ-II categories expressed in terms of CoreSC

AZ-II	CoreSC Category distribution in (%)
AIM	Object-New 27.67 % Goal 20.53% Conclusion 19.64% Result 15.18% Background 6.25% Method-New 4.46% Observation 2.67% Hypothesis 1.78% (Object-New-Advantage,Experiment) 0.89%
ANTISUPP	Result 35.29% Conclusion 29.41% Background 17.64% (Object-New, Method-New-Advantage, Observation) 5.88%
CODI	Conclusion 31.25% (Background,Result) 27.1% Method-New 6.25% Observation 4.16% (Method-New-advantage,Experiment) 2.1%
CO_GRO	Background 72.14% Motivation 10.61% Method-Old 5.3% Model 3.07% Method-Old-Advantage 2.23% Conclusion 1.67% Hypothesis 1.4% Method-Old-Disadvantage 1.11% (Object-New-Advantage,Observation,Result) 0.56% (Goal,Object-New-Advantage,Method-New-Advantage) 0.28%
FUT	Conclusion 50.9% Background 30.9% Hypothesis 7.27% Result 5.45% (Goal,Method-New,Model) 1.81%
GAP_WEAK	Background 48.88% Motivation 28.88% Method-Old-Disadvantage 11.11% (Method-Old, Conclusion) 4.44% Result 2.22%
NOV_ADV	Conclusion 49.06% Background 16.66% Method-New-Advantage 11.32% Goal 9.43% (Result,Method-Old-Advantage) 7.54% (Hypothesis,Object-New,Method-New,Motivation) 1.88%
OTHR	Background 76.63% Method-Old 8.05% Model 4.6% Experiment 3.83% (Result,Conclusion) 1.91% Hypothesis 1.14% Observation 0.77% (Motivation, Method-New-Advantage,Method-Old-Disadvantage) 0.38%
OWN_CONC	Conclusion 40.06% Result 29.25% Hypothesis 8.5% Background 7.92% Model 6.19% Observation 3.89% Method-New 1.15% Experiment 0.72% (Goal,Method-New-Advantage) 0.58% (Motivation, Method-New-Disadvantage,Object-New) 0.28% (Method-Old,Method-Old-Disadvantage) 0.14%
OWN_FAIL	Conclusion 25.8% Result 22.58% Observation 19.35% (Background,Experiment) 9.68% Model 6.45% (Method-Old-Disadvantage, Method-New-Advantage) 3.22%
OWN_MTHD	Experiment 50.84% Method-New 10.37% Background 7.78% Model 7.19% Result 5.68% Object-New 4.18% Observation 3.76% Method-Old 2.5% Goal 2.42% Hypothesis 1.76% Method-New-Advantage 1.59% Conclusion 1.34% Motivation 0.34% Method-New-Disadvantage 0.17% Object-New-Advantage 0.08%
OWN_RES	Observation 43.4% Result 42.66% Experiment 4% Conclusion 3.35% (Background,Model) 2.41% (Method-New,Method-Old) 0.52% (Goal,Object-New) 0.31%
PREV_OWN	Background 61.1% Method-Old 11.9% Experiment 7.94% Result 3.97% Motivation 3.17% (Model,Observation) 2.38% (Method-New,Conclusion) 1.59% (Hypothesis,Goal,Object-New,Method-Old-Advantage,Method-Old-Disadvantage) 0.79%
SUPPORT	Result 43.9% Conclusion 36.59% Background 14.63% Observation 4.88%
USE	Experiment 67.4% Method-Old 10.99% Method-New 7.33% Model 5.49% Background 4% (Observation,Result) 1.46% Object-New 1% Goal 0.73%
unassigned(0)	Observation 35.68% Background 19.27% Result 17.7% Conclusion 6.5% Experiment 5.47% Object-New 4.69% Method-New 3.39% Method-Old 1.56% Hypothesis 1% Motivation 0.5% (Goal,Method-New-Advantage) 0.25%

Conclusion, 14.63% Background and 4.88% Observation. ANTISUPP(*clash with other work and superiority of own*) is 35.29% Result, 29.41% Conclusion, 17.64% Background and 5.88% for each of Object-New, Method-New-Advantage and Observation. FUT(*statements about future work*) is 50.9% Conclusion, 30.9% Background, 7.27% Hypothesis and 5.45% Result. OWN_FAIL(*a solution/method/experiment in the paper that didn't work*) is 25.8% Conclusion, 22.58% Result, 19.35% Observation, 9.68% for each of Experiment and Background and 6.45% Model.

Interestingly, the categories that remain unassigned by AZ-II seem to spread across different CoreSCs, with the majority being assigned to Observation (35.68%), Background (19.27%) and Result (17.7%). This indicates that the AZ-II OWN_RES doesn't quite cover Observation and Result. Unassigned AZ-II sentences also include Conclusion (6.5%), Experiment (5.47%), Object-New (4.69%) and Method-New (3.39%).

7. Conclusion

In conclusion, the correlation between the two schemes confirms their complementary role and suggests it would be beneficial to combine the two schemes. It shows that

CoreSC categories provide a greater level of granularity when it comes to the content-related categories (e.g. (Object, Goal, Hypothesis, Motivation) vs AIM, (Method with diferent properties, Experiment, Model, Object) vs OWN_MTHD, (Observation, Result) vs OWN_RES, (Conclusion, Result, Hypothesis) vs OWN_CONC. On the other hand, AZ-II categories cover aspects of the knowledge claims that permeate across different CoreSC concepts. For example, CODI, SUPPORT, ANTISUPP, NOV_ADV illustrate the relation between the outcomes of the current work and other work, USE distinguishes between methods by other researchers used in the current work and methods introduced in the current work, whereas CO_GRO, PREV_OWN, OTHR and GAP_WEAK show the different functions of background information. The complementarity of the schemes is also illustrated in their different strengths. The highest performing categories in CoresC are Conclusion(maps mainly to OWN_CONC), Observation(maps mainly to OWN_RESand unannotated), Object(maps mainly to OWN_MTHDand AIM) and Result(maps mainly to OWN_RES) whereas for AZ-II the highest scores were obtained for USE(maps mainly to Experiment and Method-Old) and AIM(maps to Object, Goal, Conclusion). This would argue for the combination of the two schemes to make the most of their individual strenghts.

8. Applications

The CoreSC annotation scheme and the corresponding corpus were developed primarily to add semantic markup to scientific papers so as to make it easier for text mining applications to automatically access information pertaining to the content. We are currently using CoreSC annotations to train machine learning algorithms to automatically the recognise sentence based core scientific concepts in papers. We intend to use the automatically generated CoreSC annotations for extractive summarisation and intelligent querying of the papers. Other potential uses of the CoreSC annotations are information extraction, ontology population and indeed mapping to ontology codes/concepts, as the CoreSCs can indentify zones of interest were potential ontology concepts can be found (e.g. Object,Method,Observation,Experiment).

The AZ scheme on the other hand follows the relation between the current work and cited work, and is better suited to citation summaries, sentiment analysis and the extraction of information pertaining to knowledge claims. As the two schemes are complementary in their approach, a joint scheme can lead to better generated summaries, both pertaining to the content of the scientific investigation, knowledge claims and attribution to authors.
While CoreSC and AZ-II are especially designed for marking up entire papers, concurrent work is comparing three annotation schemes (including AZ and CoreSC) on annotating scientific abstracts to aid experts with cancer risk assessment.

9. References

T. Byrt, J. Bishop, and J.B Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 45(5):423–429.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Joseph L Fleiss, Jacob Cohen, and B.S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Joesph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–381.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of the IJCNLP 2008*.

A. Korhonen, L. Sun, I. Silins, and U. Stenius. 2009. The first step in the development of text mining technology for cancer risk assessment: Identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics*, 323(10).

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

M. Liakata and L.N. Soldatova. 2009. The art corpus. Technical report, Aberystwyth University.

M. Liakata, Claire Q, and S. Soldatova. 2009. Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of BioNLP-09*, pages 193–200, Boulder, Colorado.

T. McIntosh and J.R. Curran. 2009. Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10(311).

B. Medlock and T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *45th Annual Meeting of the ACL*, pages 23–30, Prague, Czech Republic.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

L.N. Soldatova and R.D. King. 2006. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3:795–803.

L.N. Soldatova and M. Liakata. 2007. An ontology methodology and cisp-the proposed core information about scientific papers. Technical Report JISC Project Report, Aberystwyth University.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, Singapore.

P. Thompson, S.A. Iqbal, J. McNaught, and S. Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(349).

V. Vincze, G. Szarvas, R. Farkas, G. Mra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.