Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics

Simone Teufel
Computer Laboratory
Cambridge University
sht25@cl.cam.ac.uk

Advaith Siddharthan
Computer Laboratory
Cambridge University
as372@cl.cam.ac.uk

Colin Batchelor
Royal Society of Chemistry
Cambridge, UK
batchelorc@rsc.org

Abstract

Argumentative Zoning (AZ) is an analysis of the argumentative and rhetorical structure of a scientific paper. It has been shown to be reliably used by independent human coders, and has proven useful for various information access tasks. Annotation experiments have however so far been restricted to one discipline, computational linguistics (CL). Here, we present a more informative AZ scheme with 15 categories in place of the original 7, and show that it can be applied to the life sciences as well as to CL. We use a domain expert to encode basic knowledge about the subject (such as terminology and domain specific rules for individual categories) as part of the annotation guidelines. Our results show that non-expert human coders can then use these guidelines to reliably annotate this scheme in two domains, chemistry and computational linguistics.

1 Introduction

Teufel et al. (1999) define the task of Argumentative Zoning (AZ) as a sentence-by-sentence classification with mutually exclusive categories from the annotation scheme given in Fig. 1. The reasoning behind the categories is inspired by the notion of a *knowledge claim* (Myers, 1992; Luukkonen, 1992): the act of writing a paper corresponds to an attempt of claiming ownership for a new piece of knowledge, which is to be integrated into the repository of scientific knowledge in the authors' field by the process of peer review and publication. In the cause of this process, the authors have to convince the reviewers that the knowledge claim of the paper is valid (Swales, 1990; Hyland, 1998). What AZ aims to model, then, are some of the relevant stages in this argument. We divide the paper into *zones*, OTHER, OWN and BACKGROUND. These are defined on the basis of who owns the knowledge claim in the corresponding segment. There are also two categories which are defined by their relationship to existing work, BASIS and CONTRAST. That means that parts of the AZ scheme are similar to citation function classification schemes from the area of citation content analysis (Garfield, 1965; Weinstock, 1971; Spiegel-Rüsing, 1977), and to automatic citation function classification (Nanba and Okumura, 1999; Garzone and Mercer, 2000; Teufel et al., 2006). The remaining categories, AIM and TEXTUAL, fulfil different rhetorical functions for the presentation of the paper. AIM points out the paper's main knowledge claim, a rhetorical move which may be repeated in the conclusion and the introduction. TEXTUAL explains the physical location of information, e.g., by giving a section overview or presenting a summary of a subsection. On the basis of human-annotated training material, AZ can be automatically classified using supervised machine learning.

Category	**Description**
AIM	Statement of research goal.
BACKGROUND	Description of generally accepted background knowledge.
BASIS	Existing KC provides basis for new KC.
CONTRAST	An existing KC is contrasted, compared, or presented as weak.
OTHER	Description of existing KC.
OWN	Description of any other aspect of new KC.
TEXTUAL	Indication of paper's textual structure.

Figure 1: AZ Annotation Scheme (Teufel et al. 1999).

Rhetorical information marking is useful for

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1493–1502,
Singapore, 6-7 August 2009.

many novel information access tasks. For instance, information retrieval can profit from rhetorical information in the form of paradigm shift statements (Chichester et al., 2005), as papers containing such statements have a high impact in an area. 75% of the ”Faculty of 1000 Biology” papers (which are chosen by experts for their special importance) contain paradigm shift sentences (Agnes Sandor, personal communication).

AZ annotation allows the construction of multi- and single document summaries which concentrate on differences and similarities to related (cited) work. AZ can also be used for search in a data base of scientific articles, in particular for enhanced citation indexing. This has been previously explored in a task-based evaluation, were users were asked to list positive and negative citations they would expect in a paper, given a short extract (Teufel, 2001). In that task, AZ-based extracts outperformed other document surrogates.

Feltrim et al. (2005) present a writing support system which analyses students' drafts of summaries for their PhD theses, performs an AZ analysis on them and critiques the rhetorical structure of the students' draft on the basis of it.

The definition of the AZ categories is based on rhetorical principles and should be decidable, in principle, without specific domain knowledge about what is discussed in detail in the paper. We present here the first evidence that AZ categories can be reliably recognised across scientific disciplines, using chemistry and computational linguistics as our model disciplines for these experiments.

The categories just introduced are abstract and depend on the annotators' interpretation of a rhetorical argument. This means that there is no guarantee that several independent annotators would annotate similarly. It is therefore crucial that all annotations at a high level of interpretation are backed up by human annotation with more than one annotator. However, annotations of citation function classification typically use only the untested annotation of a single human annotator as gold standard, who is typically the designer of a scheme (Spiegel-Rüsing, 1977; Weinstock, 1971; Nanba and Okumura, 1999; Garzone and Mercer, 2000). Teufel et al. (2006) are the only exception who test their citation function scheme using modern corpus-linguistic annotation methodology.

A study of human agreement on AZ annotation exists (Teufel et al., 1999), but this uses articles from only one discipline, namely computational linguistics. In this paper, we use a similar methodology to Teufel et al., but with data from two disciplines. The preliminary conclusion from these experiments is that annotation with chemistry papers has resulted in higher agreement than annotation with computational linguistics papers.

We extend the AZ annotation scheme to make further distinctions, as will be discussed in section 2. We also created an environment in which domain knowledge that an annotator might have about the science in a paper is systematically disregarded. We will describe how this was done in section 3, and then present the annotation experiment itself in section 4.

2 Changes to the AZ Scheme

Argumentative Zoning II (AZ-II) is a new annotation scheme, which is an elaboration of the original AZ scheme. It is presented in Fig. 2. Our annotation guidelines are 111 sides of A4 and contain a decision tree, detailed description of the semantics of the 15 categories, 75 rules for pairwise distinction of the categories and copious examples from both chemistry and computational linguistics. During guideline development, 70 chemistry papers and 20 CL papers were used, which are distinct from the ones used for annotation. It took 3 months part-time-work to prepare the guidelines for CL, and substantially less time to adapt them for chemistry. We have made them available at `www.cl.cam.ac.uk/research/nl/sciborg`.

The differences between the original AZ and AZ-II are as follows:

- Category AIM remained the same.
- Category BACKGROUND was renamed CO_GRO, or common ground.
- Category OTHER was split into other people's work (OTHR) and the authors' own previous work (PREV_OWN).
- Category BASIS was split into usage (USE) and support (SUPPORT).
- Category CONTRAST was split into neutral comparison (CODI), contradiction (ANTISUPP), and a category combining research gaps with criticism (GAP_WEAK).
- Category OWN was split into description of method (OWN_MTHD), results (OWN_RES) and conclusions (OWN_CONC), and a category which specifies recoverable errors made by the authors (OWN_FAIL).

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		

Figure 2: AZ-II Annotation Scheme.

- Category Textual was discontinued, because it is less informative than the other categories.
- Two new categories were introduced, Nov_Adv (advantages of the new knowledge claim) and Fut (declaration of limitations or future work).

Our AZ-II categories are more fine-grained than the original AZ categories. The reasons for this are twofold: To bring AZ closer to contemporary citation function schemes, and to incorporate distinctions recently found useful by other researchers. For instance, Chichester et al. (2005) argue that Antisupp is particularly important. The finer grain in AZ-II has been accomplished purely by splitting existing AZ categories; hence, the coarser AZ categories are recoverable (with the exception of the Textual category). Annotation examples are given in the appendix.

As in AZ, citations are an important but not necessarily decisive cue for a sentence to belong to a particular zone. The guidelines mention citations as one factor in deciding whether a knowledge claim holds, and citations occur in several examples, so it is likely that the presence of citations would have influenced annotators in their decision.

Of the changes, the distinction which is likely to have the greatest impact on the annotation is the split of Own according to the stage of the authors' problem solving process – into methods, results, conclusion or local failure. In most life sciences, descriptions of research as a problem solving process are a dominant phenomenon, whereby problem-solving descriptions can be of differing length and embeddedness. For instance, in synthetic chemistry, the starting compound for the main synthesis in the paper may first have to be synthesised itself (if it is not commercially available, for instance). In that case, arriving at the compound is an intermediate, smaller problem-solving process which enables the larger problem-solving process that represents the new KC.

The original AZ scheme didn't mark the distinction, possibly because it is not as easily observable in CL as it is in the life sciences, and because problem-solving stages were not part of the main analytic interest of AZ, which focused on how scientific argumentation is related to descriptions of own and other work. Also, neither of the traditional AZ applications (summarisation or citation indexing) had any direct use for the subdivided categories. But in the life sciences, there are applications which would make use of such a subdivision. For instance, in chemistry there is a niche for search applications which guide searchers directly to the method and/or result sections in papers. Specifically, the Own_Fail category is motivated by the failure–and–recovery search. In text, Own_Failmarks cases where the authors helpfully mention in passing steps which were found not to work during a long synthetic procedure (often the 'total synthesis' of a compound which is found in nature). Such cases happen frequently, and are generally followed by a 'recovery' statement which explains how the problem can be avoided. Another possible application that calls for a subdivision is Feltrim et al.'s

(2005) rhetorical writing system for novice writers. It trains novices in writing rhetorically well-formed abstracts and therefore must have a way of distinguishing, for instance, between methods and results.

Note that several of the applications based on AZ and AZ-II in general rely on the rare categories much more than they rely on the more frequent categories. OWN_FAIL is an example of a rare but important category, and so is AIM, which is central to summarisation applications. The comparative and contrastive categories CODI ANTISUPP and GAP_WEAK, on the other hand, are particularly useful to citation-based search applications.

Other AZ-like schemes for scientific discourse created for the biomedical domain (Mizuta and Collier, 2004) and for computer science (Feltrim et al., 2005) also made the decision to subdivide OWN, in similar ways to how we propose here. The current work, however, is the first experimental proof that humans can make this distinction – and the others encoded in AZ-II – reliably, and in two quite distinct disciplines.

3 Discipline-Independent Non-Expert Annotation

An important principle of AZ is that its categories can be decided without domain knowledge. This rule is anchored in the guidelines: when choosing a category, no reasoning about the scientific facts is allowed. The avoidance of domain-knowledge has its motivation in a strategy for a hypothetical automatic text-understanding system for unrestricted texts. Given the state of the art in text processing and knowledge representation, text understanding systems should in our opinion use general, rhetorical, and logical aspects of the text, rather than attempting to recognise or represent the scientific knowledge contained in the text. What the human annotation – the gold standard – should then do is to simulate the best possible output that such a system could theoretically create.

Annotators may use only general, rhetorical or linguistic knowledge; knowledge which is shared by all proficient speakers of a language. The guidelines spell out what is meant by these general principles. For instance, one can use lexical and syntactic parallelism in a text to infer that the authors were setting up a comparison between themselves and some other approach.

There is, however, a problem with annotator expertise and with the exact implementation of the "no domain knowledge" principle. This problem does not become apparent until one starts working with disciplines where at least some of the annotators or guideline developers are not domain experts (chemistry, in our case). Domain experts naturally use scientific knowledge and inference when they make annotation decisions. It would be unrealistic to expect them to be able to disregard their domain knowledge simply because they were *instructed* to do so. Additionally, when all annotators/scheme developers are domain experts, it is hard to even notice the cases where they "accidentally" use domain knowledge during annotation. We therefore artificially created a situation where all annotators are "semi-informed non-experts", which forces them to comply with the principle, namely by the following rules:

Justification: Annotators have to justify all annotation decisions by pointing to some text-based evidence, and by giving the section heading in the guidelines that describes the particular reason for assigning the category. General discipline-specific knowledge an annotator may happen to have is excluded as justification. Annotators' justifications have to be typed into the annotation tool and are open to challenge during the training phase. Much of the allowable justification comes in the form of general and linguistic principles, e.g., an explicit cue phrase, the title, or the structural similarity of textual strings. For instance, annotators are allowed to infer that process-VPs in the title are likely to be the contribution (knowledge of the actual concrete contribution of a paper is a requirement for annotation of AIM).

Discipline-specific Generics: The guidelines contain a section with high-level facts about the general research practices in the discipline. These generics constitute the only *scientific* knowledge which is acceptable as a justification, and are aimed to help non-expert annotators recognise how a paper might relate to already established scientific knowledge, so that they will be able to avoid common mistakes about the knowledge claim status of a certain fact. For instance, the better they are able to distinguish what is commonly known from what is newly claimed by the authors, the more consistent their annotation will be.

Annotation with expert-trained non-expert annotators means that a domain expert must be available initially, during the development of the annotation scheme and the guidelines, either as a co-developer or as an informant. The domain expert's job is to describe scientific knowledge in that domain in a general way, in as far as it is necessary for the scheme's distinctions, and to write the domain-specific rules for the individual categories, including the choice of example sentences. This means that the guidelines are split into a domain-general and a domain-specific part.

The discipline-specific generics in chemistry come in the form of a "chemistry primer", a 10-page collection of high-level scientific domain knowledge. It contains: a glossary of words a non-chemist would not have heard about or would not necessarily recognise as chemical terminology; a list of possible types of experiments performed in chemistry; a list of commonly used machinery; a list of non-obvious negative characterisations of experiments and compounds ("sluggish", "inert"); and a list of possible types of knowledge claims. For instance, in chemistry each chemical substance mentioned can have in principle a knowledge claim associated with its discovery or invention – with the exception of water, rock salt, the metals known in prehistory and a few others. If a compound or process is however considered to be so commonly used that it is in the "general domain" (e.g., "the Stern–Volmer equation" or "the Grignard reaction"), it is no longer associated with somebody's knowledge claim, and as a result its usage is not to be marked with category USE.

Descriptions of individual categories can have domain-specific subsections, as well as the general ones. For instance, if the text states that the authors could not replicate a published result, the guidelines describe the cases when this is the authors' fault (OWN_FAIL) in contrast to the cases where this is an indirect accusation of the previous experiment (ANTISUPP).

Another potentially unclear distinction is between results (OWN_RES) and conclusions (OWN_CONC). The difference is defined on the basis of how much reasoning is necessary to be able to make the statement concerned. If all the authors did was to read a measurement off an instrument, the label OWN_RES applies. Reasoning points to OWN_CONC; it is sometimes linguistically marked ("therefore", "we conclude", "this means that"), but in many cases, domain knowledge may be required to decide whether reasoning was necessary to make a certain statement. Possible OWN_RES statements, according to the chemistry primer, include: statements of simple numerical result; descriptions of graphs; descriptions of atoms' positions in three-dimensional space; statements of trends, unless a reason for these results is given; comparisons of results of more than one experiment, unless a reason for these results is given.

The chemistry primer also lists phenomena which in a typical experiment would be read off chemical machinery (e.g., "Stark effect"). This list gives the non-expert annotator an objective criterion to answer the question how likely it is that a certain statement by the authors was arrived at by inference. We also found that our list of phenomena which can be read off machinery, which was compiled from the first 30 papers, generalised well to the other 40 papers considered.

The chemistry primer is not an attempt to summarise all methods and experimentation types in chemistry; this would be impossible to do, certainly in a few pages. Rather, it tries to answer many of the high-level questions a non-expert would have to an expert, in the framework of AZ.

This methodology allows to hire expert and non-expert annotators and bring them in line with each other. We believe it could be expanded relatively easily into many other disciplines, using domain experts which create similar primers for genetics, experimental physics, cell biology, but re-using the bulk of the guidelines.

4 Annotation Experiments

The annotators were the co-developers of the annotation scheme and the authors of this paper. Whereas all three annotators have good background knowledge in CL, the largest difference between them concerns their expertise in chemistry: Annotator A is a PhD-level chemist, Annotator B has two years' of undergraduate training in chemistry and can therefore be considered a chemical semi-expert, and Annotator C has no specialised chemistry knowledge.

As agreement measure we choose the Kappa coefficient κ (Fleiss, 1971; Siegel and Castellan, 1988), the agreement measure predominantly used in natural language processing research (Carletta, 1996). κ corrects raw agreement $P(A)$ for agreement by chance $P(E)$:

$$\kappa = \frac{P(A)-P(E)}{1-P(E)}$$

No matter how many items or annotators, or how the categories are distributed, $\kappa = 0$ when there is no agreement other than what would be expected by chance, and $\kappa = 1$ when agreement is perfect. If two annotators agree *less* than expected by chance, κ can also be negative. Chance agreement $P(E)$ is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators. All work done here is reported in terms of Fleiss' κ. [1] κ is also designed to abstract over the number of annotators as its formula relies on the proportion of expected vs. observed *pairwise* agreements possible in a pool. That is, κ for k annotators will be an average of the values of κ taking all possible m-tuples of annotators from the annotator pool (with $m < k$). As a side effect of its definition of random agreement, κ treats agreement in a rare category as more surprising, and rewards such agreement more than an agreement in a frequent category. This is a desirable property, because we are more interested in the performance of the rare rhetorical categories than we are in the performance of the more frequent zone categories.

Category	**Chem**	**CL**	**Category**	**Chem**	**CL**
OWN_MTHD	25.4	55.6	SUPPORT	1.5	0.7
OWN_RES	24.0	5.6	GAP_WEAK	1.1	1.0
OWN_CONC	15.1	10.7	FUT	1.0	1.4
OTHR	8.3	10.0	NOV_ADV	1.0	0.8
USE	7.9	2.7	CODI	0.8	1.2
CO_GRO	6.7	5.7	OWN_FAIL	0.8	0.1
PREV_OWN	3.4	1.7	ANTISUPP	0.5	0.6
AIM	2.3	1.8			

Figure 3: Frequency of AZ-II Categories (in %).

4.1 Data

For chemistry, 30 random-sampled papers from journals published between 2004 and 2007 by the Royal Society of Chemistry were used for annotation[2]. The papers cover all areas of chemistry and some areas close to chemistry, such as climate modelling, process engineering, and a double-blind medical trial. The data used for the experiment contains a total of 3745 sentences.

For computational linguistics, 9 papers were annotated, with a total of 1629 sentences. The papers were published between 1998 and 2001 at ACL, EACL or EMNLP conferences, and were taken from the Computation and Language archive. Both chemistry and CL papers were automatically sentence-split, with manual correction of errors; acknowledgement sections were disregarded. A web-based annotation tool was used for guideline definition and for annotation.

Our choice of which data sets to use was effected by the relative length of papers more than by the journal/conference distinction. Average article length between chemistry journal articles (3650 words/paper) and CL conference articles (4219 words/paper) is comparable, so conference articles in CL seem a much better choice for comparative work than journal publications, which are often very long in CL. Additionally, conferences have a high profile in CL, and we found the conference publications to be of high editorial quality. We are nevertheless interested in the structure of longer journal articles, and plan to investigate CL journals in the future.

The annotations were done using a web-based annotation tool. Every sentence is assigned a category. No communication between the annotators was allowed.

4.2 Results

The inter-annotator agreement for chemistry was κ = 0.71 (N=3745,n=15,k=3). For CL, the inter-annotator agreement was κ = 0.65 (N=1629,n=15,k=3). For comparison, the inter-annotator agreement for the original, CL-specific AZ with 7 categories was κ = 0.71 (N=3420,n=7,k=3). Given the subjective nature of the task and the fact that AZ-II introduces additional distinctions, the AZ-II agreement can be considered acceptable for CL and relatively high for chemistry. Additionally, chemistry annotation used one non-expert annotator, who had no chemistry-specific domain knowledge apart from that in the chemistry primer.

The distribution of categories for the two disciplines is given in Fig. 3. As expected, there is a large discrepancy in frequency between the (rare) rhetorical categories and the (much more frequent) zone categories OWN_MTHD, OWN_RES,

[1] Artstein and Poesio (2008) observe that there are several version of κ which differ in how $P(E)$ is calculated. In particular, Fleiss' (1971) κ calculates $P(E)$ as the average observed distribution of all annotators, whereas Cohen's (1960) κ calculates $P(E)$ only on the basis of the other annotator(s).

[2] 100 papers across a spread of disciplines from the January 2004 issues of the RSC were selected blindly (but with an attempt to cover most areas of chemistry). 30 out of these were random sampled for annotation; the rest were used for annotation development.

OWN_CONC, OTHR and CO_GRO. For supervised learning, too few examples of any category can be a problem. There are methods which attempt to reduce the annotation effort by using a trained classifier to suggest possible cases to a human. However, the classifier can only find examples similar to the ones that have already been manually classified, when the real problem is a recall-problem, i.e., the challenge is to find more new examples in the multitude of possible sentences. To solve this in a fundamentally sound way, there seems to be no other way than to annotate more texts, at the cost of more human effort.

If we consider the differences across disciplines, the most striking ones concern the distribution of OWN_MTHD, which is more than twice as common in CL (56% v. 25%), and OWN_RES, which is far more common in chemistry overall (24% v 5.6%). Usage of other people's knowledge claims or materials also seems to be more common in chemistry, or at least more explicitly expressed (7.9% vs 2.7%). With respect to the shorter, rarer categories, there is a marked difference in OWN_FAIL (0.1% in CL, but 0.8% in chemistry[3] and SUPPORT, which is more common in chemistry (1.5% vs 0.7%). However, this effect is not present for ANTISUPP (contradiction of results), the "reverse" category to SUPPORT, (0.6% in CL vs 0.5% in chemistry).

As far as the chemistry annotation is concerned, it is interesting to find out whether Annotator A was influenced during annotation by domain knowledge which Annotator C did not have, and Annotator B had to a lower degree[4]. We therefore calculated pairwise agreement, which was $\kappa_{\mathrm{AC}} = 0.66$, $\kappa_{\mathrm{BC}} = 0.73$ and $\kappa_{\mathrm{AB}} = 0.73$ (all: N=3745,n=15,k=2). That means that the largest disagreements were between the non-expert (C) and the expert (A), though the differences are modest. This might point to the fact that Annotators A and B might have used a certain amount of domain-knowledge which the chemistry primer in the guidelines does not yet, but should, cover.

In an attempt to determine how well categories are defined, we first consider the binary distinction between zone categories (OWN_MTHD, OWN_RES, OWN_CONC, OWN_FAIL, OTHR, PREV_OWN and CO_GRO) and rhetorical categories (the other 8). This shows an inter-annotator agreement of κ_{binary} = 0.78 (N=3745, n=2, k=3) for chemistry and κ_{binary} = 0.65 (N=1629, n=2, k=3) for CL, indicating that annotators find it relatively easy (chemistry) or at least not more difficult than the overall distinction (CL) to distinguish these two types of categories. We next perform Krippendorff's (1980) category distinctions (Fig. 4). Here, all categories apart from the one diagnosed are collapsed, and what is reported is the difference of inter-annotator agreement when compared to the overall distinctiveness (κ=0.71 for chemistry, κ=0.65 for CL). Where the difference is positive, the annotators could distinguish the given category better than they could distinguish all categories, and where they are negative, correspondingly worse.[5]

The results confirm that categories USE, AIM, OWN_MTHD, OWN_RES and FUT are particularly well distinguished in both disciplines. This is a positive result, as these categories are important for several types of searches. In these cases the guidelines seem to fully suffice for their description, but then again good performance of AIM, FUT and USE is not that surprising, as they are signalled clearly by linguistic and non-linguistic cues. However, there are three categories with particularly low distinguishability in both disciplines: ANTISUPP, OWN_FAIL and PREV_OWN. As ANTISUPP and OWN_FAIL are crucial for the envisaged downstream tasks, the problems with their definition should be identified and fixed. We are in the process of systematically troubleshooting the guidelines for those categories.

The table also shows that category definition has discipline-specific problems. For instance, we believe that the fact that distinctiveness for OWN_FAIL is so bad for CL must be due to the fact that we only encountered very few potential OWN_FAIL cases in this domain. The definition of the categories SUPPORT and NOV_ADV also seem to be substantially more confusing for CL than for chemistry. However, CODI is a category which shows average distinctiveness for CL, but much worse distinctiveness for chemistry. We believe this is due to the fact that comparisons of

[3] These are not large differences in absolute terms – 55 items identified as OWN_FAIL by at least one annotator in chemistry, vs. 7 such items in CL, the relative difference is large and confirms that in chemistry papers, particularly descriptions of synthesis procedures, OWN_FAIL cases appear relatively frequently.

[4] This question does not arise in the case of CL, as all annotators can be considered experts in this respect.

[5] All κ values for chemistry were measured with N=3745, n=2, k=3; for CL with N=1629, n=2, k=3.

methods and approaches are more common in CL and are clearly expressed, whereas in chemistry the objects that are involved in comparisons are more varied and at a lower grade of abstraction (e.g., compounds, properties of compounds, coefficients, etc.), which obviously has a negative effect on the distinctiveness of this category.

Category	Chem	CL	Category	Chem	CL
USE	+0.12	+0.00	NOV_ADV	-0.07	-0.23
AIM	+0.09	+0.08	OWN_CONC	-0.08	-0.13
OWN_MTHD	+0.05	+0.05	GAP_WEAK	-0.08	-0.16
OWN_RES	+0.02	+0.04	PREV_OWN	-0.11	-0.15
FUT	+0.01	+0.06	OWN_FAIL	-0.19	-0.43
CO_GRO	-0.01	-0.03	ANTISUPP	-0.35	-0.32
SUPPORT	-0.04	-0.12	CODI	-0.36	+0.00
OTHR	-0.06	+0.07			

Figure 4: Krippendorff's Diagnostics for Category Distinction (κ, relative to Overall Distinctiveness).

We also provide a direct comparison of our annotation results with those from the original AZ scheme. Comparisons between two similar annotation schemes can be made by collapsing those categories in each scheme which are not distinguished in the other scheme. Such a comparison can of course only ever approximate the smallest common denominator between two schemes.

The AZ-II categories were collapsed into a set of six categories that closely resemble AZ categories, as described in section 2 (with OWN simulated by the union of OWN_FAIL, OWN_MTHD, OWN_RES, OWN_CONC, FUT, and NOV_ADV). This created a 6-category AZ annotation.

As TEXTUAL is not marked up in AZ-II, the original AZ annotation was also collapsed, by incorporating TEXTUAL examples into OWN. The two 6-pronged AZ-annotations are now more directly comparable. Inter-annotator agreement for the collapsed AZ-II showed κ = 0.75 (N=3745, n=6, k=3). This compares favourably to the collapsed AZ's agreement of κ = 0.71 (N=3420, n=6, k=3); but when comparing the raw numerical results one should consider that different data from different disciplines is used (chemistry in AZ-II, CL in AZ).

These results should be interpreted as a positive result for the domain-independence of AZ, and also for the feasibility of using trained non-experts as annotators. The additional work that went into the guidelines has produced annotation of a high consistency, even though AZ-II provides more distinctions (15 categories vs. 7 in AZ).

There is also the faint possibility that discourse annotation of chemistry is intrinsically easier than discourse annotation of CL, *because* it is a more established discipline and not despite of it. For instance, it is likely that the problem-solving categories OWN_FAIL, OWN_MTHD, OWN_RES and OWN_CONC are easier to describe in a discipline with an established methodology (such as chemistry), than they are in a younger, developing discipline such as computational linguistics.

5 Conclusion

Argumentative Zoning is an analysis of the rhetorical progression of the scientific argument in a paper. In this paper, we have made the following contributions to this analysis:

- We have presented a more informative scheme, which additionally recognises the structure of an experiment in terms of problem solving (method – results – conclusions) and makes more fine-grained distinctions in some of the sentiment-inspired relational categories (e.g., criticism and comparisons to other approaches).
- We introduced an annotation methodology which attempts to systematically exclude the use of annotators' extraneous domain knowledge from the annotation.
- We have experimentally shown that human coders can independently annotate this new AZ scheme in two distinct disciplines. Our results show inter-annotator agreements of κ=0.65 and κ=0.71 for computational linguistics and chemistry, respectively.

Overall, the outcome of this work indicates that the phenomena described in AZ can be defined in a domain-independent way. The experiment also tested how realistic the "expert-trained non-expert" approach to domain-knowledge free annotation is. The fact that the agreement between three annotators (an expert, a semi-expert, and a non-expert) is acceptable overall vindicate our task definition as domain-knowledge free (using the tools of justification and domain-specific generic knowledge). However, the agreements involving the semi-expert are higher than the agreement between expert and non-expert. This probably means that the chemistry generics were not fully adequate to ensure that the non-expert understood enough of the chemistry to achieve the highest-possible agreement.

The automation of AZ-annotation is underway. This requires adaptation of the high-level features used in AZ (Teufel and Moens, 2002) to chemistry. We are also preparing an annotation experiment with naive annotators. Another research avenue is the expansion of the guidelines to other disciplines such as bio-medicine, and to longer journal articles, e.g., in computational linguistics.

6 Acknowledgements

This work was funded by EPSRC project Sciborg (EP/C010035/1).

Appendix: Annotation Examples[6]

Aim *We now describe in this paper a synthetic route for the functionalisation of the framework of mesoporous organosilica by free phosphine oxide ligands, which can act as a template for the introduction of lanthanide ions.* (b514878b)

Aim *The aim of this paper is to examine the role that training plays in the tagging process …* (9410012)

Nov_Adv *Moreover, the simplicity and ease of application of the electrochemical method [...] should also be emphasised and makes it an interesting and valuable synthetic tool.* (b513402a)

Nov_Adv *Other than the economic factor, an important advantage of combining morphological analysis and error detection/correction is the way the lexical tree associated with the analysis can be used to determine correction possibilities.* (9504024)

Co_Gro *A wide range of organosulfur compounds are biologically active and some find commercial application as fungicides and bactericides*[1−4]*.* (b514441h)

Co_Gro *It has often been stated that discourse is an inherently collaborative process …* (9504007)

Othr *In their system, antibody immobilized on a solid substrate reacts with antigen, which binds with another antibody labelled with peroxidase.* (b313094k)

Othr *But in Moortgat's mixed system all the different resource management modes of the different systems are left intact in the combination and can be exploited in different parts of the grammar.* (9605016)

Prev_Own *As a program aimed at the applications of imines*(2a,g,5) *we have studied the formation of carbanions from imines and their subsequent reactions.* (b200198e)

Prev_Own *Earlier work of the author (Feldweg 1993; Feldweg 1995a) within the framework of a project on corpus based development of lexical knowledge bases (ELWIS) has produced LIKELY …* (9502038)

Own_Mthd *In order for it to be useful for our purposes, the following extensions must be made:* (0102021)

Own_Mthd *On the other hand, a tertiary amide can be an excellent linking functional group.* (b201987f)

Own_Fail *Initial attempts to improve the dehydration of* **4** *via chemical or thermal means were unsuccessful; similarly, attempts to couple the chlorosilane (Me3Si)2 (Me2ClSi)CH with Ag2O failed.* (b510692c)

Own_Fail *When the ABL algorithms try to learn with two completely distinct sentences, nothing can be learned.* (0104006)

Own_Res *While the acid 1a readily coupled to the olefin, the corresponding boronic ester was surprisingly inert under the reaction conditions.* (b311492a)

Own_Res *All the curves have a generally upward trend but always lie far below backoff (51% error rate).* (0001012)

Own_Conc *It is unlikely that every VOC emit ted by plants serves an ecological or physiological role …* (b507589k)

Own_Conc *Unless grammar size takes on proportionately much more significance for such longer inputs, which seems implausible, it appears that in fact the major problems do not lie in the area of grammar size, but in input length.* (9405033)

Gap_Weak *Various methods of preparation have been developed, but they often suffer from low yield and tedious separation.*[16,17,28,31] (b200888m)

Gap_Weak *Here, we will produce experimental evidence suggesting that this simple model leads to serious overestimates of system error rates…* (9407009)

CoDi *However, the measured values of the dielectric constant (ϵ = 310) are lower than the values reported by Ganguli and coworkers*(21) *for BSTO pellets sintered at 1100* deg*C …* (b506578j)

CoDi *Unlike most research in pragmatics that focuses on certain types of presuppositions or implicatures, we provide a global framework in which one can express all these types of pragmatic inferences.* (9504017)

Support *This is in line with the findings of Martin and Illas for inorganic solids* (84,85)*.* (b515732c)

Support *Work similar to that described here has been carried out by Merialdo (1994), with broadly similar conclusions.* (9410012)

Use *The diamine* **10** *was prepared following a previously published procedure*(4d)*.* (b110865b)

Use *We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988).* (9504007)

Fut *Our further efforts are directed towards the above goal,…and overcoming limitations pertaining to the electron-poor arylboronic acids.* (b311492a)

Fut *An important area for future research is to develop principled methods for identifying distinct speaker strategies pertaining to how they signal segments.* (9505025)

Antisupp *Although purification of 8b to a de of 95percent has been reported elsewhere*[31]*, in our hands it was always obtained as a mixture of the two [EQN]-diastereomers.* (b310767a)

Antisupp *This result challenges the claims of recent discourse theories (Grosz and Sidner 1986, Reichman 1985) which argue for a the close relation between cue words and discourse structure.* (9504006)

[6] Corpus examples are taken from our chemistry and CL data sets; indicated by their respective file numbers.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Christine Chichester, Frdrique Lisacek, Aaron Kaplan, and Agnes Sandor. 2005. Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. In *Proceedings of First International Symposium on Semantic Mining in Biomedicine*.

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

V. Feltrim, Simone Teufel, Gracas Nunes, and S. Alusio. 2005. Argumentative zoning applied to critiquing novices' scientific abstracts. In Janyce Wiebe James G. Shanahan, Yan Qu, editor, *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–245. Springer, Dordrecht, The Netherlands.

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

Eugene Garfield. 1965. Can citation indexing be automated? In M. et al. Stevens, editor, *Statistical Association Methods for Mechanical Documentation (NBS Misc. Pub. 269)*. National Bureau of Standards, Washington.

Mark Garzone and Robert E. Mercer. 2000. Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the CSCI/SCEIO (AI-2000)*, pages 337–346.

Ken Hyland. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, 30(4):437–455.

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

Terttu Luukkonen. 1992. Is scientists' publishing behaviour reward-seeking? *Scientometrics*, 24:297–319.

Yoko Mizuta and Nigel Collier. 2004. An annotation scheme for rhetorical analysis of biology articles. In *Proceedings of LREC'2004*.

Greg Myers. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of the XXth International Joint Conference on Artificial Intelligence (IJCAI-99)*, pages 926–931.

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

Ina Spiegel-Rüsing. 1977. Bibliometric and content analysis. *Social Studies of Science*, 7:97–113.

John Swales, 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*, pages 110–176. Cambridge University Press, Cambridge, UK.

Simone Teufel and Marc Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–446.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pages 110–117, Bergen, Norway.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of EMNLP-06*.

Simone Teufel. 2001. Task-based evaluation of summary quality: Describing relationships between scientific papers. In *Proceedings of NAACL-01 Workshop "Automatic Text Summarization"*, Pittsburgh, PA.

Melvin Weinstock. 1971. Citation indexes. In *Encyclopedia of Library and Information Science*, volume 5, pages 16–40. Dekker, New York, NY.