A Framework for Annotating 'Related Works' to Support Feedback to Novice Writers

Casey, Arlene

ACL
2019

Casey , A , Webber , B & Glowacka , D 2019 , A Framework for Annotating 'Related Works' to Support Feedback to Novice Writers . in 13TH LINGUISTIC ANNOTATION WORKSHOP (LAW XIII) . ACL , pp. 90-99 , 13th Linguistic Annotation Workshop (LAW) , Florence , Italy , 01/08/2019 . https://doi.org/10.18653/v1/W19-4011

http://hdl.handle.net/10138/317333
https://doi.org/10.18653/v1/W19-4011

A Framework for Annotating 'Related Works' to Support Feedback to Novice Writers

Arlene Casey
School of Informatics
University of Edinburgh
Edinburgh, UK
a.j.casey@sms.ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh, UK
bonnie@inf.ed.ac.uk

Dorota Głowacka
Dept. of Computer Science
University of Helsinki
Helsinki, Finland
glowacka@cs.helsinki.fi

Abstract

Understanding what is expected of academic writing can be difficult for novice writers to assimilate, and recent years have seen several automated tools become available to support academic writing. Our work presents a framework for annotating features of the *Related Work* section of academic writing, that supports writer feedback.

1 Introduction

Learning the skill of academic writing is critical for post-graduate students to be successful, yet many struggle to master the standard of quality expected of them (Aitchison et al., 2012; Paltridge and Starfield, 2007). Beyond the surface characteristics of grammar and spelling, students must grasp aspects of style and content structure expected within their discipline. Automated recognition of content features in academic writing has become a popular approach to assist students in recent years. Previous work has focused on identifying rhetoric intentions, such as those described by Swales (1981) that can be found in an Introduction (Cotos and Pendar, 2016; Anthony and V. Lashkia, 2003) or in PhD summaries (Feltrim et al., 2006). Other approaches have focused on identifying argument components and relations and how these relate to essay scores (Ghosh et al., 2016). The one aspect that these approaches have in common is the need for annotated data based on task-orientated annotation schemes. Our focus is on building an annotation schema which can help writers recognise appropriate intentions in writing their *Related Work* section, and indicate when these are missing.

Annotating intention in academic writing is challenging as the language and author intentions differ across the typical sections found in a paper (Introduction, Methods, Results, Discussion) and within disciplines (Hyland, 2015). We focus on one section of scientific text that has, for the most part, been ignored in the past — the *Related Work* section.

Currently no annotation schema specifically focuses on *Related Work*. There are schemas that capture some, but not all, elements of intentions we seek, such as those that consider citation function (Teufel et al., 2006a; Angrosh et al., 2012) or argument zones reflecting rhetoric intentions (Teufel, 1999; Teufel et al., 2009). However, these are designed for different purposes, such as understanding citation relations, summarisation or information extraction (e.g gene relations, knowledge claims). Thus, they also have labels that are irrelevant to *Related Work*, e.g. 'Conclusion', which may make the annotation task more difficult. Since previous work has shown that annotation schemes benefit from being designed for their specific goal (Guo et al., 2010), we propose a specific annotation framework to support automated writing feedback on *Related Work*.

This paper describes our framework for annotating the discourse of *Related Work* in such a way that it supports feedback on writing. The framework reflects qualities that both theory and experiments have shown to be important. We discuss how these qualities have motivated our design along with those existing schemes that are most closely related to ours. We report results that show reliable annotation for this framework. Future work will investigate the degree to which such annotation can be automated.

2 Background

Our aim is to help authors recognise rhetorical intentions that are present in their writing and highlight those that are missing, using these intentions

Proceedings of the 13th Linguistic Annotation Workshop, pages 90–99
Florence, Italy, August 1, 2019.

to form our feedback to writers. Argument structures are key in allowing an author to convey and provide a persuasive message, which forms the author intention. Swales (1981) was one of the first to recognise author intentions, calling them *rhetorical moves*, a strategy employed by a writer to strengthen the persuasive appeal or stage of an argument. This section discusses work on developing annotation schemes related to identifying rhetoric intentions in scientific publications and writing analytic tools. We highlight some of the challenges others have found when working with intentions in scientific publications and how this relates to our goal of writing feedback. Section 3 provides more detail how schemas directly map to our annotation design. Subsequent sections describe the dataset we have used (Section 4), the annotation process and results (Sections 5–6), and our plans to develop the work further (Section 7).

2.1 Understanding Author Intent in Scientific Discourse

Argument Zoning (AZ) (Teufel, 1999) was one of the first author intention schemas to provide reliability studies of their annotations and to fully automate these. AZ marks zones that identify knowledge claims indicating who these knowledge claims belong to, in addition to providing categories for relationships between the authors or existing works. Teufel et al. (2009) extended the AZ schema from 7 to 15 categories. This extension allowed the authors to then apply their schema to the domain of life sciences in addition to their original domain of Computational Linguistics. The AZ scheme has also been successfully adapted in other domains, e.g. biology (Mizuta and Collier, 2004). The requirement to adapt the schema to new domains supports the idea that different styles of writing across domains may influence recognising intention in writing and our choice to focus on only one domain.

Whilst the AZ scheme has proven very successful, it has been applied to capturing intentions across entire documents. The schema was designed to support tasks of summarisation and to improve information access. For a section such as *Related Work*, which is rarely used in summarisation or information access, this means that its meaningful author intentions may be labelled too generically to be useful or not at all. Nevertheless, AZ has been shown to be successful for feedback on abstracts and summaries of PhD's (Feltrim et al., 2006). As one of the intended goals of the AZ schema is summarisation, it is not surprising that the schema works well for this type of writing feedback.

Understanding the motivation or function of a citation can help determine an author's intention (Teufel et al., 2006a). Work is not meant to be cited simply because it is on the same topic as the citing work. Rather, cited works should be ones that have implications for the author's study (Maxwell, 2006). The development of citation schemas, with corresponding annotations, has been a subject of research for several decades (Weinstock, 1971; Oppenheim and Renn, 1978; Teufel et al., 2006a; Angrosh et al., 2012). However, many of the early citation studies are based on small samples, and do not include reliability studies as annotation is done by the author only. Such weak annotation methodology could lead to unforeseen difficulties when it comes to practical implementation of these schemas. There is agreement that determining the relationship of the cited work to that of the author(s) can be difficult, and that this subjective nature makes it hard to operationalise (Teufel, 1999; Swales, 1990). Often context is linguistically unmarked, which can make judgements about the relationship of the cited work more difficult to make (Teufel, 1999). We believe that novice writers struggle to provide citations that go beyond lists or brief description, and this leads to what Teufel calls "linguistically unmarked context". We also believe the reader's experience has a role to play in interpretation, with experts in the field not requiring as many linguistic clues to relevance as a novice reader may require.

Our work differs from most other citation frameworks in that determining whether the author made the citation relevant in context to their own work is more important than the recognition of the citation function. For example, highlighting that there is a gap in a cited work is not our primary focus. We want to capture that a gap is highlighted but also further ensure it is made relevant to the authors' own work e.g. they state what they do that is different to fill the gap. Identifying neutral or linguistically unmarked citations is important as they indicate an opportunity for feedback that the writing may need revision to clarify relevance.

Some work specifically looks at developing

annotation frameworks which are more directly linked to the Toulmin model of argumentation (Toulmin, 2003) to represent argument structures in a research article. These annotation schemas represent arguments as claims and premises with some including relations of support and attack (Stab and Gurevych, 2014). Whilst this structure has been shown to work well in a persuasive essay scenario, it would not support the types of intentions discussed in the next section that are relevant to *Related Work*.

2.2 Writing Analytics Tools

Using rhetoric intentions to provide writers with feedback has been successful in academic writing. Mover (Anthony and V. Lashkia, 2003), Research Writer Tutor (RWT) (Cotos and Pendar, 2016) and ACAWriter (Abel, 2018) are three tools based on Swales CARS model (Swales, 1990). The first two tools carry out annotation based on their interpretation of the CARS model — the first on the *Abstract* and the second on the *Introduction*. Unfortunately, little information is provided on the annotation process. There are indications that the RWT is intended to be used as a University tool, so perhaps propriety concerns are behind restricting the availability of information or annotated datasets. However, as the CARS model is designed for the *Introduction*, this makes it likely any schemas would be only partially relevant to identifying content expected in *Related Work*.

Whilst previous works motivate our approach, no other work provides a match for the fine-grained author intentions that allow informative writing feedback for *Related Work*. It is known that annotation schemas benefit from being task-orientated (Guo et al., 2010). Hence, we see a need to develop an annotation schema for recognising author intentions in *Related Work* sections that meet the goal of writer feedback.

3 Annotation Schema for *Related Work*

Domain Disciplines differ in their writing conventions for academic papers. As a result, linguistic constructs and content can differ across disciplines (Hyland, 2015). Not all disciplines have a specific *Related Work* section – some include literature material in the *Introduction* or disperse it throughout other sections. Due to these challenges, we focus on the discipline of Computational Linguistics, where *Related Work* sections are more readily found.

Annotation unit We have chosen the sentence as our unit of annotation. Many other works mentioned in the background section, such as those based on AZ, use sentence as an annotation unit. We acknowledge that using a sentence could introduce challenges – for example, a given sentence could potentially serve two functions that may be better captured at clause level. For our purposes of providing feedback, we believe the sentence as an annotation unit will be the most meaningful. One reason for this is that in the next stage of our work (providing feedback), we will need to look at several sentences together to determine relevance, as citation relevance has been shown to require to look beyond just the citing sentence (Teufel et al., 2006a).

3.1 The Annotation Schema

We first consider what qualities should be present in the *Related Work* section of a paper in Computational Linguistics and then we discuss how we map these into our annotation schema.

Identifying Qualities in a Related Work Section We base our *Related Work* qualities on key tasks that Kamler and Thomson (2006, p. 28) indicate a survey of related work should accomplish.

- ***Background*** This information has an important goal of helping the author to locate their work in the field, showing they understand their field and its history through indicating seminal works and other relevant research fields. They may provide some evidence through citation to what they are saying.

- ***Cited Works*** From more generally identifying the field, the author should demonstrate specifically (i) which works, methods or ideas are most pertinent to their work; (ii) how these works have influenced and motivated what they do; and (iii) if and how the current work builds on or uses these methods.

- ***Gap*** In addition to demonstrating what works are most pertinent to their work, the author should also make clear what the gap is, what areas or applications have not yet been addressed in existing work. This can be done when citing specific work or it could be indicated as a gap in the field when discussing background.

Literature Quality	Sentence Label	Description
Background	BG-DESC-NE	Description of the state of the field, describing/listing known methods or common knowledge. No evidence i.e. citation is not included
	BG-DESC-EP	Description of the state of the field, describing/listing known methods or common knowledge. Evidence provided i.e.citation included
	BG-EVAL-P	Author highlights a positive aspect in the field
Cited Work	CW-DESC	Describes cited work, this could be specific details, or very high level details or nothing more than a reference for further information
	CW-COMP	Cited work compared to another cited work
	CW-EVAL-P	Positive aspect highlighted of cited work
	A-CW-BUILD	Author's work uses/builds on (adapts/modifies) cited work
	A-CW-SIM	Author's work is similar to cited work
Gap	CW-EVAL-SC	Shortcoming, problem or gap about the cited work is highlighted
	BG-EVAL-SC	Author highlights a shortcoming, problem or gap in the field
Author Contribution	A-DIFF	Author states their work is different with no detail
	A-DESC	Author describes their work with no linguistic marking to other's work or being different
	A-GAP	Author specifically says they address a gap or highlights the novelty of their work
	A-CW-DIFF	Author's highlights how their work is different to cited work
Additional Labels	OTHER	Sentence does not fit under any other label
	OCR	Sentence has OCR problems and annotator cannot understand
	TEXT	Sentence provides information about what will be discussed in the next section

Table 1: Annotation Labels

- ***Contribution*** Having exposed a gap, the author should indicate their contribution to address this gap and highlight what makes their work different or novel.

3.2 Mapping Qualities to the Annotation Schema

Looking just at label names, it can seem like our labels (Table 1) are direct replications of other schemas. However, on closer inspection of how authors' apply these labels, we often find discrepancies that would not work for our purpose. Table 2 provides a discussion of comparisons and similarities of our label schema to those that are most closely related (Fisas et al., 2015, 2016; Teufel, 1999; Teufel et al., 2006b; Angrosh et al., 2012; Teufel et al., 2009). One contributing factor as to why existing labels do not adequately support our goals is that they are designed to look across the whole of a document. As a result, they seek either very general or much finer grained labelling than we require. For example, Fisas et al. (2016) distinguishes between an author using methods, using data or using tools from another cited work. This finer grained approach is not relevant or needed to provide feedback in a *Related Work* section, we only need to know that the author used the cited work.

3.3 Qualities and their corresponding labels

Background These types of sentences describe the state of the field, common knowledge, or describe/list known methods. We ask our annotators to identify two types of background sentences — (i) with citations i.e. evidence provided – BG-DESC-EP and these citations are not part of the syntax of the sentence. (ii) Background sentences without evidence i.e. no citations – BG-DESC-NE. Part of the reason for this distinction is that novice writers make a limited use of citation types (Thompson and Tribble, 2001). We also include a background label that relates to when an author says something positive or highlights a strength in the field/general – BG-EVAL-P.

Cited Works To provide informative feedback, we need to establish the relevance of a cited work to the author's work or if this cited work is perfunctory in nature. Firstly, we provide a label that accounts for description of a cited work – CW-DESC. Our other labels account for contrasting the author's work to cited work saying: (i) it is similar – A-CW-SIM; (ii) the author uses/builds on or adapts/modifies the cited work – A-CW-BUILD. Teufel et al. (2006b) describes a category CoCoXY that contrasts two pieces of cited work, and highlights that this is often not annotated in the literature as most works put comparisons to author's work and a cited work together. This dis-

Quality/Our Labels	Related Works	Comparison
Background		
BG-DESC-NE BG-DESC-EP	(Teufel, 1999) (Liakata et al., 2012) (Angrosh et al., 2012) (Fisas et al., 2015)	All the related works use a label of 'Background' but they do not distinguish between those that have citation evidence or not. There are some discrepancies in what these capture to ours for e.g in Angrosh this is used for *sentences that provide background or introduction*. Fisas in addition to sentences that state common ground includes sentences of previous related work in their background category. The reason for their more general approach could be attributed to these other works capturing labels across the whole article.
BG-EVAL-P	-	We did not find evidence of other works looking for strengths in background sentences.
Cited Works		
CW-DESC	(Teufel et al., 2006b) (Angrosh et al., 2012) (Fisas et al., 2016)	Teufel and Fisas have a category 'Neutral' which is directly related to our category of CW-DESC. These are used like our label for descriptions of a cited work. Fisas differs slightly in that they also include in this category *references for more information* or *comments on common practices* which we would put in one of our 'Background' sentence labels. Teufel also allows this label to be used for an *unlisted citation function* or *not enough evidence to put in any other category*. In our case these would go into the *OTHER* label. Angrosh provides two labels 'RWD_CS' – *a sentence describing a citation occurring in that sentence*, 'RWD' – *a sentences describing a related work where the citation does not occur in that sentence*. Our one label covers both Angrosh's labels.
A-CW-SIM	(Teufel et al., 2006b) (Fisas et al., 2016)	Both Fisas with a label of 'Comparison-similarity' and Teufel with a label of 'PSim' have categories that label sentences with *authors work is similar to the cited work*.
A-CW-BUILD	(Teufel et al., 2006b) (Fisas et al., 2016)	Fisas and Tuefel have labels which align with our category of A-CW-BUILD. However, they break this into finer detail than we feel is necessary for our goal. Fisas has four labels for using another cited work: 'Use-method, 'Use-Data', 'Use-Tool', 'Use-other' and three labels for authors work based on a cited work, 'Basis-previous own work', 'Basis Others work', 'Basis -future work'. Teufel has three labels: 'PBas', *uses cited work as basis*, 'PUse', *author uses tools/algorithms/data/definition*, 'PModi', *author adapts or modifies tools/algorithms/data*. This finer grained approach supports the goal of these authors as they look across a whole document but is not necessary for our goal of writer feedback.
CW-COMP	(Teufel et al., 2006b)	Teufel includes a category CoCoXY which contrasts two pieces of cited work as our sentence label does.
CW-EVAL-P	(Angrosh et al., 2012) (Fisas et al., 2016)	Angrosh has two labels that represent what we capture here RWS_CS and RWS. The first of these labels mentions a positive (strength) in a citation sentence and in the second a positive (strength) is mentioned but the citation is not present in that sentence. Fisas also has this label 'CRITICISM-Strength'.
Gap		
CW-EVAL-SC	(Fisas et al., 2016)(Teufel et al., 2006b) (Angrosh et al., 2012)	Our evaluation category for cited works relates directly to Tuefel's category of 'Weak' - *weakness of cited approach* and Fisas's 'Criticism-weakness'. Angrosh labels this as 'RWSC' - *sentence noting the shortcomings in the related work citation.*
BG-EVAL-SC	(Teufel et al., 2009)	Tuefel's work is the only evidence of where we can find a similarity to our label of a shortcoming in the field although her label 'GAP_WEAK' - *lack of solution in field, problem with other solutions* covers a shortcoming in both the field and a cited work.
Contribution		
A-GAP	(Fisas et al., 2016)(Teufel et al., 2009)	This has similarities to Fisas's 'Novelties', although their label is not exclusive to the author's approach and could include other cited work. Teufel's category of 'NOV-ADV' is for sentences claiming a novelty or advantage of the author's own approach
A-CW-DIFF	(Fisas et al., 2016)	Our category of author and cited work comparison, A-CW-D, directly relates to the category of Fisas of 'Comparison-difference'.
A-DESC	(Teufel et al., 2009)	We could not find a schema that labels sentences just as author description. Other works such as Tuefel have several labels which in part fall under this category such as :'OWN_MTHD, OWN_FAIL,OWN_RES,OWN_CONC, AIM'. These are very specific and likely not to occur very often in a Related Work.
TXT	(Teufel, 1999)	In her original AZ schema Teufel includes a label of TEXT that is the same as our label.

Table 2: Label Schema Comparison

tinction of comparing two works rather than the author's work and a cited work is important for recognising how an author makes citations relevant. Therefore, we incorporate this category into our schema as – CW-COMP. Additionally, we include a label for an author highlighting a positive or strength of a cited work – C-EVAL-P.

Gap Locating a gap in academic writing often takes the approach of highlighting weaknesses or areas not addressed in others' work or in the field in general. We also want to identify when a gap or shortcoming is highlighted in the field in general. We add two categories: (i) BG-EVAL-SC for a background sentence highlighting a gap/weakness in the field; (ii) CW-EVAL-SC, where an author highlights a shortcoming, problem or gap about a specific cited work.

Author Contribution Here, we want to capture if the author specifically identifies how they will address a gap. This is done by authors when they specifically say their work is novel, new or describe how they address a gap with the label – A-GAP. Our label A-CW-DIFF applies when an author compares their work directly with a cited work, saying it is different and how it is different. We also capture where an author describes their own work – A-DESC. This type of description may not linguistically identify that the author has made a contribution but the explanation may describe this novelty or difference to others' work. Here, it could be expected that a reader's experience may allow them to interpret this as a contribution but we instruct our annotators only to mark it as contribution if it is linguistically marked. The identification of this type of sentence is less common in other schemas.

3.4 Learning from Pilot Annotations

Initially, a preliminary annotation study was conducted that highlighted a problem when considering author differences. There were many occurrences of an author sentence which just indicated "our work is different", giving no details why or how. The annotators pointed out that these were not very informative sentences and quite different to when the author actually provides details of why their work is different. The extra label, A-DIFF, was added to account for this.

In addition, there were some sentences which had OCR problems, so a category was created for this, along with a category for TEXT. TEXT indicates where an author says "In the next section we will discuss". This type of category was in the original AZ schema, but we thought it unlikely to arise in a *Related Work* section. However, it was highlighted in the pilot annotations. A category of OTHER was also added as there were some sentences the annotators could not assign to a label.

4 Dataset

Initial experiments were carried out on a pre-annotated dataset (Schäfer et al., 2012) consisting of 266 published scientific papers from the ACL anthology (Bird et al., 2008). The dataset was extracted from PDF by commercial OCR software, sentence-tokenised and then manually annotated, using MMAX2 (Müller and Strube, 2006). Papers were annotated for co-reference to cited papers and to the authors' own work. All the papers were 6 to 8 pages long. This is important, as short-conference papers (4 pages) would have considerably shorter *Related Work* sections. Initially, we processed the full data set, and then only those papers with *Related Work* sections were extracted. This resulted in a data set of 113 papers. Our final dataset comprised of the 95 *Related Work* sections that remained after we removed papers with OCR problems.

Authors do not always signal the relevance of a paper in its citing sentence: often it will come in the next or subsequent sentence. Although we are only assigning a label to a sentence, in future work it will be necessary to look at all sentences related to a citation to determine what feedback to give. This was our reason for choosing a dataset that was already marked for co-references to citations.

5 Annotation Process

5.1 Annotators

Both our annotators were PhD students in Computational Linguistics, in the final stages of their degree programs. Because knowledge possessed by researchers in a field can (in some instances) be used to overcome a lack of explicit linguistic marking, PhD students were preferable over domain experts in terms of bringing some, but not a lot of, knowledge to the task. This fact was acknowledged by Teufel et al. (2009) who instruct their annotators to only use rhetorical linguistic knowledge but point out how difficult it is for do-

main experts not to use their knowledge when annotating.

One annotator annotated the whole corpus and the other just over half the corpus (i.e., 53 *Related Work* sections).

5.2 Annotator Task

The *Related Work* sections were given to each annotator in an Excel file. Each row represented a sentence, with fields corresponding to document id, sentence id, the original sentence, and the sentence with citation and co-references marked. In the following field, the annotator entered a label from the pre-populated list provided. The final field was for comments, or for indicating any annotations they were not sure about.

5.3 Annotator Support

The annotators were given 9 pages of guidelines which contained examples and suggested workflow to decide on an annotation label. Initially, the annotators met to discuss the guidelines and ensure their understanding. They trained on the same 10 *Related Work* sections and compared their results discussing any difference.

6 Annotation Results

6.1 Corpus Analysis

The annotated corpus includes 95 *Related Works* sections and a total of 1,806 sentences. Double annotation was done for 53 *Related Works* and 955 sentences. The size of our dataset is comparable to others who have studied scientific publications in annotation. Fisas et al. (2015) studied a corpus of 40 documents, Teufel et al. (2009) studied 90 papers, Feltrim et al. (2006) 52 abstracts, and Anthony and V. Lashkia (2003) 100 abstracts.

Our results focus on the part of the corpus that double annotation was completed on to show the inter-annotator agreement and highlight the challenges.

6.2 Measuring Inter Annotator Agreement

We use Cohen's *k* (Cohen, 1960) to measure our annotator agreement, correcting for chance agreement. The formula is:

$$K = \frac{P_o - P_e}{1 - P_e}, \quad (1)$$

where P_o is observed and P_e is expected agreement. The range of Kappa can be between -1 and 1, where 0 means agreement is only expected by chance. A value of 0.8 is considered good agreement.

	CW-DIFF	DESC	DIFF	GAP
CW-DIFF	69	8	5	7
DESC	1	44	0	1
DIFF	-	-	2	-
GAP	5	6	2	23

Table 3: Author Label Agreement Matrix. The letter A (Author) at the beginning of each entry was omitted for the sake of clarity.

Kappa measures are widely used in annotation agreement in scientific publications in schemes that have been successful in automated classification based on their annotations (Teufel et al., 2009; Liakata et al., 2012; Fisas et al., 2016). In general, work on author intentions that uses Kappa agreement reports agreement in a range of 0.65-0.78 (Teufel et al., 2006a; Fisas et al., 2015; Teufel et al., 2009) with Liakata et al. (2012) being much lower at 0.55.

Teufel et al. (2009) points out that Kappa treats agreement in rare categories as surprising and rewards these more than frequent categories. Although she sees this as an advantage because scientific publications often have these rare categories, others see this as misleading and criticise that chance-corrected measures do this when applied to unbalanced data-sets. Hence, others often report raw agreement (Kirschner et al., 2015). Our data does have rare categories and so we report the raw agreement in addition to the Kappa agreement.

6.2.1 Inter-annotator Agreement

The inter-annotator agreement (IAA) was 0.77 (N = 955, n = 53, K = 2). Raw agreement was 80.1%. These results demonstrate good agreement and are comparable to similar studies mentioned earlier.

Out of the 955 sentences doubly annotated, the annotators agreed on 764. Based on the agreed sentences, the most frequent category was CW-DESC (32.5%), followed by the background categories BG-DESC-EP (12.2%) and BG-DESC-EP (10.9%). Following this were the author categories A-CW-DIFF (9%), A-CW-SIM (8.8%), A-DESC (5.8%) and A-GAP (3%). In the next section, we discuss some of the difficulties the annotators had with A-COMP-DIFF versus A-GAP/A-DESC. CW-EVAL-SC was surprisingly infrequent

	BG-DESC-EP	BG-DESC-NE	CW-DESC
BG-DESC-EP	83	10	16
BG-DESC-NE	2	93	6
CW-DESC	6	5	248

Table 4: Cited Work and Background Label Agreement Matrix

at 3.9% and CW-COMP at 2.23%. OCR and OTHER were both at 1.3%. All the remaining categories constituted less then 1% of sentences and interestingly all of these had good agreement. OCR will not occur in our writing feedback as we will not be processing text from PDF. However, OTHER or TEXT could happen, although these were rare categories with TEXT having 13 sentences in agreement and OTHER 10 sentences in agreement. TEXT was almost in perfect agreement, while OTHER was used more frequently by one annotator.

6.3 Sources of Disagreement

There were two main sources of disagreement between the annotators: one was in agreeing the labels about the author's work, and the other was in distinguishing between background sentences and those that pertained to specific citations.

In particular, the annotators noticed that when an author spoke about how their work was different to someone else's, they often broke this down over several sentences. The guidelines instructed the annotators to only mark what was linguistically indicated but they were unsure if this meant in the text in general or in that particular sentence. This led to annotators disagreeing on A-COMP-DIFF and A-GAP/A-DESC, as can be seen in Table 3. Our annotation guidelines need to be reviewed with some very specific examples that incorporate these scenarios with clear instructions on how to take linguistic markings into account. This will be a challenge for automated classification of our labels and writing feedback. We need to consider carefully how this lexical information can be captured.

In disagreement about background sentences compared to citation sentences, seen in Table 4, one annotator highlighted that some sentences talked about two specific citations and they labelled these as BG-DESC-EP, while the other annotator labelled them as CW-DESC. After discussion, it was suggested that including examples of this kind in the annotation guidelines would have helped.

Annotators also noted that a sentence may belong to two labels. For example, a sentence may say something positive about a cited work but then highlight a shortcoming. In the guidelines we instruct the annotator to choose the author based labels over cited work labels in this instance. We acknowledged in choosing the sentence as the annotating unit this could occur, and we think this will prove challenging in automating the labelling.

There were two *Related Work* sections that included references to systems by their names, e.g. Moses or U-SVM. The annotators struggled with both of these as they were only given the *Related Work*. If they had the full paper, they thought they would better ascertain if the author was referring to something that was their own work or another person's. One annotator questioned whether these types of *Related Work* were more likely to come at the end of a paper once a reader was familiar with these terms. Neither annotator thought the guidelines could be updated as in this instance it would have been better to have access to the full paper. Again, this is going to be a challenging area for any automated system, especially if it only takes a submission of the *Related Work* section into account. The system will have no way of knowing if phrases of this kind relate to the author's work. It also raises a point that although we have chosen one discipline to work with, *Related Work* sections can still be written in different styles. Prior to this comment, we had not considered if order within a document impacted the style of the *Related Work*. However, it should still fulfil the qualities expected.

6.4 Annotating the Remaining Sentences

Following a discussion between the annotators on labels that were not in agreement, some changes were made. A small number of the disagreements were genuine mistakes with an annotator selecting the wrong label but most were about the differences in A-COMP-DIFF versus A-GAP/A-DESC, and between CW-DESC and the Background categories. This resulted in an increase in agreement to 0.85 and raw agreement to 87.3%. Part of the reason for this discussion and alignment was to ensure that the annotator who had completed the full corpus was confident about their decisions. They reviewed the remaining sentences following the discussion. The labels from the annotator who

completed all sentences will be used as the standard for the full corpus to develop our automated system in the future.

7 Conclusions and Future Work

We have developed a new annotation schema designed to capture author intentions in *Related Work* sections. Our annotation scheme focuses on qualities that should be present in a *Related Work* section and that will support writing feedback. Our schema has 14 categories that will be used in feedback. We report good agreement in our annotation, which is comparable to other annotation experiments within our field. Our experiments help us to refine our annotation guidelines for any future annotation activities and make us aware of challenges we may encounter when trying to automate the classification of the labels within our schema for feedback.

In future work we plan to use our annotated corpus in supervised machine learning to automate the classification of our labels. Work is currently underway to determine features that will best represent the schema labels, taking into account the challenges our annotators raised. This classification model will be an important part of our automated writing system. However, this classifier will treat sentences as individual components, and we need to put these sentences into context to provide meaningful feedback. Future work will involve experiments to investigate how context can be derived from combining the individual labels to provide feedback that adequately reflects the writing.

References

Kitto Kirsty Knight Simon Buckingham Shum Simon Abel, Sophie. 2018. Designing personalised, automated feedback to develop students research writing skills. In *Proceedings of 2018 Australasian Society for Computers in Learning in Tertiary Education*, pages 15–24.

Claire Aitchison, Janice Catterall, Pauline Ross, and Shelley Burgin. 2012. 'Tough love and tears': learning doctoral writing in the sciences. *Higher Education Research & Development*, 31(4):435–447.

M A Angrosh, Stephen Cranefield, and Nigel Stanger. 2012. Context identification of sentences in research articles: Towards developing intelligent tools for the research community. *Natural Language Engineering*, 19(04):481–515.

Laurence Anthony and George V. Lashkia. 2003. Mover: A Machine Learning Tool to Assist in the Reading and Writing of Technical Papers. *Professional Communication, IEEE Transactions on*, 46:185 – 193.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *LREC 2008*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Elena Cotos and Nick Pendar. 2016. Discourse classification into rhetorical functions for awe feedback. *calico journal*, 33(1):92–116.

Valéria D Feltrim, Simone Teufel, Maria Graças V das Nunes, and Sandra M Aluísio. 2006. Argumentative zoning applied to critiquing novices scientific abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–246. Springer.

Beatríz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *LREC 2016*.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. *Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes*. Association for Computational Linguistics, Uppsala, Sweden.

Ken Hyland. 2015. Genre, discipline and identity. *Journal of English for Academic Purposes*, 19(C):32–43.

Barbara Kamler and Pat Thomson. 2006. *Helping doctoral students write: Pedagogies for supervision*. Routledge.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. pages 1–11.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Joseph A. Maxwell. 2006. Literature Reviews of, and for, Educational Research: A Commentary on Boote and Beile's "Scholars Before Researchers". *Educational Researcher*, 35(9):28–31.

Yoko Mizuta and Nigel Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 29–35. Association for Computational Linguistics.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Charles Oppenheim and Susan P Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):225–231.

Brian Paltridge and Sue Starfield. 2007. *Thesis and Dissertation Writing in a Second Language*. Routledge.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology. In *Proceedings of COLING 2012: Posters*, pages 1059–1070, Mumbai, India. The COLING 2012 Organizing Committee.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

J. Swales. 1981. Aspects of article introductions. *Language Studies Unit*.

J.M. Swales. 1990. *Genre Analysis: English in academic and research settings*. Cambridge University Press.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006a. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87, Sydney, Australia. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006b. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Paul Thompson and Chris Tribble. 2001. Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning Technology*, 5(3):91 – 105.

Stephen E Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Melvin Weinstock. 1971. Citation indexes. encyclopedia of library and information science. volume 5. eds. a. kent & h. lancour.