An annotation scheme for citation function

Simone Teufel **Advaith Siddharthan** **Dan Tidhar**
Natural Language and Information Processing Group
Computer Laboratory
Cambridge University, CB3 0FD, UK
{Simone.Teufel,Advaith.Siddharthan,Dan.Tidhar}@cl.cam.ac.uk

Abstract

We study the interplay of the discourse structure of a scientific argument with formal citations. One subproblem of this is to classify academic citations in scientific articles according to their rhetorical function, e.g., as a rival approach, as a part of the solution, or as a flawed approach that justifies the current research. Here, we introduce our annotation scheme with 12 categories, and present an agreement study.

1 Scientific writing, discourse structure and citations

In recent years, there has been increasing interest in applying natural language processing technologies to scientific literature. The overwhelmingly large number of papers published in fields like biology, genetics and chemistry each year means that researchers need tools for information access (extraction, retrieval, summarization, question answering etc). There is also increased interest in automatic citation indexing, e.g., the highly successful search tools Google Scholar and CiteSeer (Giles et al., 1998).[1] This general interest in improving access to scientific articles fits well with research on discourse structure, as knowledge about the overall structure and goal of papers can guide better information access.

Shum (1998) argues that experienced researchers are often interested in relations between articles. They need to know if a certain article criticises another and what the criticism is, or if the current work is based on that prior work. This type of information is hard to come by with current search technology. Neither the author's abstract, nor raw citation counts help users in assessing the relation between articles. And even though CiteSeer shows a text snippet around the physical location for searchers to peruse, there is no guarantee that the text snippet provides enough information for the searcher to infer the relation. In fact, studies from our annotated corpus (Teufel, 1999), show that 69% of the 600 sentences stating contrast with other work and 21% of the 246 sentences stating research continuation with other work do not contain the corresponding citation; the citation is found in preceding sentences (i.e., the sentence expressing the contrast or continuation would be outside the CiteSeer snippet). We present here an approach which uses the classification of citations to help provide relational information across papers.

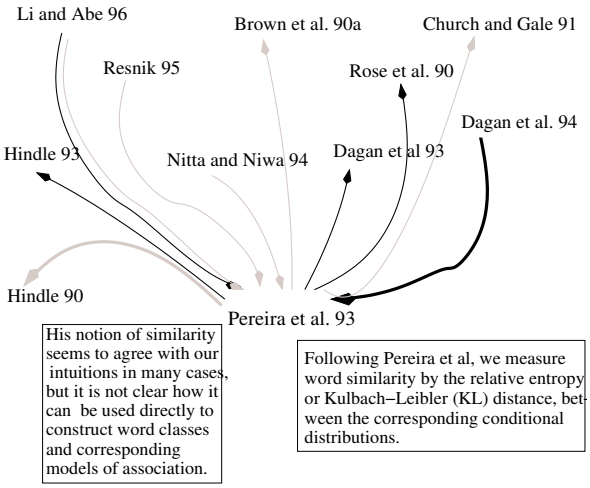

Figure 1: A rhetorical citation map

Citations play a central role in the process of writing a paper. Swales (1990) argues that scientific writing follows a general rhetorical argumentation structure: researchers must justify that their paper makes a contribution to the knowledge in their discipline. Several argumentation steps are required to make this justification work, e.g., the statement of their specific goal in the paper (Myers, 1992). Importantly, the authors also must relate their current work to previous research, and acknowledge previous knowledge claims; this is done with a formal citation, and with language connecting the citation to the argument, e.g., statements of usage of other people's approaches (often near textual segments in the paper where these approaches are described), and statements of contrast with them (particularly in the discussion or related work sections). We argue that the automatic recognition of citation function is interesting for two reasons: a) it serves to build better citation indexers and b) in the long run, it will help constrain interpretations of the overall argumentative structure of a scientific paper.

Being able to interpret the rhetorical status of a citation at a glance would add considerable value to citation indexes, as shown in Fig. 1. Here differences and similarities are shown between the example paper (*Pereira et al., 1993*) and the papers it cites, as well as

[1] CiteSeer automatically citation-indexes all scientific articles reached by a web-crawler, making them available to searchers via authors or keywords in the title.

Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pages 80–87,
Sydney, July 2006.

the papers that cite it. *Contrastive* links are shown in grey – links to rival papers and papers the current paper contrasts itself to. *Continuative* links are shown in black – links to papers that are taken as starting point of the current research, or as part of the methodology of the current paper. The most important textual sentence about each citation could be extracted and displayed. For instance, we see which aspect of *Hindle (1990)* the *Pereira et al.* paper criticises, and in which way *Pereira et al.*'s work was used by *Dagan et al.* (1994).

We present an annotation scheme for citations, based on empirical work in content citation analysis, which fits into this general framework of scientific argument structure. It consists of 12 categories, which allow us to mark the relationships of the current paper with the cited work. Each citation is labelled with exactly one category. The following top-level four-way distinction applies:

- Weakness: Authors point out a weakness in cited work
- Contrast: Authors make contrast/comparison with cited work (4 categories)
- Positive: Authors agree with/make use of/show compatibility or similarity with cited work (6 categories), and
- Neutral: Function of citation is either neutral, or weakly signalled, or different from the three functions stated above.

We first turn to the point of how to classify citation function in a robust way. Later in this paper, we will report results for a human annotation experiment with three annotators.

1.	Cited source is mentioned in the introduction or discussion as part of the history and state of the art of the research question under investigation.
2.	Cited source is the specific point of departure for the research question investigated.
3.	Cited source contains the concepts, definitions, interpretations used (and pertaining to the discipline of the citing article).
4.	Cited source contains the data (pertaining to the discipline of the citing article) which are used sporadically in the article.
5.	Cited source contains the data (pertaining to the discipline of the citing particle) which are used for comparative purposes, in tables and statistics.
6.	Cited source contains data and material (from other disciplines than citing article) which is used sporadically in the citing text, in tables or statistics.
7.	Cited source contains the method used.
8.	Cited source substantiated a statement or assumption, or points to further information.
9.	Cited source is positively evaluated.
10.	Cited source is negatively evaluated.
11.	Results of citing article prove, verify, substantiate the data or interpretation of cited source.
12.	Results of citing article disprove, put into question the data as interpretation of cited source.
13.	Results of citing article furnish a new interpretation/explanation to the data of the cited source.

Figure 2: Spiegel-Rüsing's (1977) Categories for Citation Motivations

2 Annotation schemes for citations

In the field of library sciences (more specifically, the field of Content Citation Analysis), the use of information from citations above and beyond simple citation counting has received considerable attention. Bibliometric measures assesses the quality of a researcher's output, in a purely quantitative manner, by counting how many papers cite a given paper (White, 2004; Luukkonen, 1992) or by more sophisticated measures like the h-index (Hirsch, 2005). But not all citations are alike. Researchers in content citation analysis have long stated that the classification of motivations is a central element in understanding the relevance of the paper in the field. Bonzi (1982), for example, points out that *negational* citations, while pointing to the fact that a given work has been *noticed* in a field, do not mean that that work is *received well*, and Ziman (1968) states that many citations are done out of "politeness" (towards powerful rival approaches), "policy" (by name-dropping and argument by authority) or "piety" (towards one's friends, collaborators and superiors). Researchers also often follow the custom of citing some particular early, basic paper, which gives the foundation of their current subject ("paying homage to pioneers"). Many classification schemes for citation functions have been developed (Weinstock, 1971; Swales, 1990; Oppenheim and Renn, 1978; Frost, 1979; Chubin and Moitra, 1975), inter alia. Based on such annotation schemes and hand-analyzed data, different influences on citation behaviour can be determined, but annotation in this field is usually done manually on small samples of text by the author, and not confirmed by reliability studies. As one of the earliest such studies, Moravcsik and Murugesan (1975) divide citations in running text into four dimensions: conceptual or operational use (i.e., use of theory vs. use of technical method); evolutionary or juxtapositional (i.e., own work is based on the cited work vs. own work is an alternative to it); organic or perfunctory (i.e., work is crucially needed for understanding of citing article or just a general acknowledgement); and finally confirmative vs. negational (i.e., is the correctness of the findings disputed?). They found, for example, that 40% of the citations were perfunctory, which casts further doubt on the citation-counting approach.

Other content citation analysis research which is rel-

evant to our work concentrates on relating textual spans to authors' descriptions of other work. For example, in O'Connor's (1982) experiment, *citing statements* (one or more sentences referring to other researchers' work) were identified manually. The main problem encountered in that work is the fact that many instances of citation context are linguistically unmarked. Our data confirms this: articles often contain large segments, particularly in the central parts, which describe other people's research in a fairly neutral way. We would thus expect many citations to be neutral (i.e., not to carry any function relating to the argumentation per se).

Many of the distinctions typically made in content citation analysis are immaterial to the task considered here as they are too sociologically orientated, and can thus be difficult to operationalise without deep knowledge of the field and its participants (Swales, 1986). In particular, citations for general reference (background material, homage to pioneers) are not part of our analytic interest here, and so are citations "in passing", which are only marginally related to the argumentation of the overall paper (Ziman, 1968).

Spiegel-Rüsing's (1977) scheme (Fig. 2) is an example of a scheme which is easier to operationalise than most. In her scheme, more than one category can apply to a citation; for instance positive and negative evaluation (category 9 and 10) can be cross-classified with other categories. Out of 2309 citations examined, 80% substantiated statements (category 8), 6% discussed history or state of the art of the research area (category 1) and 5% cited comparative data (category 5).

Category	Description
Weak	Weakness of cited approach
CoCoGM	Contrast/Comparison in Goals or Methods (neutral)
CoCoR0	Contrast/Comparison in Results (neutral)
CoCo-	Unfavourable Contrast/Comparison (current work is better than cited work)
CoCoXY	Contrast between 2 cited methods
PBas	author uses cited work as starting point
PUse	author uses tools/algorithms/data
PModi	author adapts or modifies tools/algorithms/data
PMot	this citation is positive about approach or problem addressed (used to motivate work in current paper)
PSim	author's work and cited work are similar
PSup	author's work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories or unlisted citation function

Figure 3: Our annotation scheme for citation function

Our scheme (given in Fig. 3) is an adaptation of the scheme in Fig. 2, which we arrived at after an analysis of a corpus of scientific articles in computational linguistics. We tried to redefine the categories such that they should be reasonably reliably annotatable; at the same time, they should be informative for the application we have in mind. A third criterion is that they should have some (theoretical) relation to the particular discourse structure we work with (Teufel, 1999).

Our categories are as follows: One category (Weak) is reserved for weakness of previous research, if it is addressed by the authors (cf. Spiegel-Rüsing's categories 10, 12, possibly 13). The next three categories describe comparisons or contrasts between own and other work (cf. Spiegel-Rüsing's category 5). The difference between them concerns whether the comparison is between methods/goals (CoCoGM) or results (CoCoR0). These two categories are for comparisons without explicit value judgements. We use a different category (CoCo-) when the authors claim their approach is better than the cited work.

Our interest in differences and similarities between approaches stems from one possible application we have in mind (the rhetorical citation search tool). We do not only consider differences stated between the current work and other work, but we also mark citations if they are explicitly compared and contrasted with other work (not the current paper). This is expressed in category CoCoXY. It is a category not typically considered in the literature, but it is related to the other contrastive categories, and useful to us because we think it can be exploited for search of differences and rival approaches.

The next set of categories we propose concerns positive sentiment expressed towards a citation, or a statement that the other work is actively used in the current work (which is the ultimate praise). Like Spiegel-Rüsing, we are interested in use of data and methods (her categories 4, 5, 6, 7), but we cluster different usages together and instead differentiate unchanged use (PUse) from use with adaptations (PModi). Work which is stated as the explicit starting point or intellectual ancestry is marked with our category PBas (her category 2). If a claim in the literature is used to strengthen the authors' argument, this is expressed in her category 8, and vice versa, category 11. We collapse these two in our category PSup. We use two categories she does not have definitions for, namely similarity of (aspect of) approach to other approach (PSim), and motivation of approach used or problem addressed (PMot). We found evidence for prototypical use of these citation functions in our texts. However, we found little evidence for her categories 12 or 13 (disproval or new interpretation of claims in cited literature), and we decided against a "state-of-the-art" category (her category 1), which would have been in conflict with our PMot definition in many cases.

Our fourteenth category, Neut, bundles truly neutral descriptions of other researchers' approaches with all those cases where the textual evidence for a citation function was not enough to warrant annotation of that category, and all other functions for which our scheme did not provide a specific category. As stated above, we do in fact expect many of our citations to be neutral.

Citation function is hard to annotate because it in principle requires interpretation of author intentions (what could the author's intention have been in choosing a certain citation?). Typical results of earlier citation function studies are that the sociological aspect of citing is not to be underestimated. One of our most fundamental ideas for annotation is to only mark explicitly signalled citation functions. Our guidelines explicitly state that a general linguistic phrase such as "better" or "used by us" must be present, in order to increase objectivity in finding citation function. Annotators are encouraged to point to textual evidence they have for assigning a particular function (and are asked to type the source of this evidence into the annotation tool for each citation). Categories are defined in terms of certain objective types of statements (e.g., there are 7 cases for `PMot`). Annotators can use general text interpretation principles when assigning the categories, but are not allowed to use in-depth knowledge of the field or of the authors.

There are other problematic aspects of the annotation. Some concern the fact that authors do not always state their purpose clearly. For instance, several earlier studies found that negational citations are rare (Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977); MacRoberts and MacRoberts (1984) argue that the reason for this is that they are potentially politically dangerous, and that the authors go through lengths to diffuse the impact of negative references, hiding a negative point behind insincere praise, or diffusing the thrust of criticism with perfunctory remarks. In our data we found ample evidence of this effect, illustrated by the following example:

> *Hidden Markov Models (HMMs)* **(Huang et al. 1990)** *offer a powerful statistical approach to this problem, though it is unclear how they could be used to recognise the units of interest to phonologists.* (9410022, S-24)[2]

It is also sometimes extremely hard to distinguish usage of a method from statements of similarity between a method and the own method. This happens in cases where authors do not want to admit they are using somebody else's method:

> *The same test was used in* **Abney and Light (1999)**. (0008020, S-151)

> *Unification of indices proceeds in the same manner as unification of all other typed feature structures* **(Carpenter 1992)**. (0008023, S-87)

In this case, our annotators had to choose between categories `PSim` and `PUse`.

It can also be hard to distinguish between continuation of somebody's research (i.e., taking somebody's research as starting point, as intellectual ancestry, i.e. `PBas`) and simply using it (`PUse`). In principle, one would hope that annotation of all usage/positive categories (starting with `P`), if clustered together, should result in higher agreement (as they are similar, and as the resulting scheme has fewer distinctions). We would expect this to be the case in general, but as always, cases exist where a conflict between a contrast (`CoCo`) and a change to a method (`PModi`) occur:

> *In contrast to* **McCarthy, Kay and Kiraz**, *we combine the three components into a single projection.* (0006044, S-182)

The markable units in our scheme are a) all full citations (as recognized by our automatic citation processor on our corpus), and b) all names of authors of cited papers anywhere in running text outside of a formal citation context (i.e., without date). Our citation processor recognizes these latter names after parsing the citation list an marks them up. This is unusual in comparison to other citation indexers, but we believe these names function as important referents comparable in importance to formal citations. In principle, one could go even further as there are many other linguistic expressions by which the authors could refer to other people's work: pronouns, abbreviations such as "Mueller and Sag (1990), henceforth M & S", and names of approaches or theories which are associated with particular authors. If we could mark all of these up automatically (which is not technically possible), annotation would become less difficult to decide, but technical difficulty prevent us from recognizing these other cases automatically. As a result, in these contexts it is impossible to annotate citation function directly on the referent, which sometimes causes problems. Because this means that annotators have to consider non-local context, one markable may have different competing contexts with different potential citation functions, and problems about which context is "stronger" may occur. We have rules that context is to be constrained to the paragraph boundary, but for some categories paper-wide information is required (e.g., for `PMot`, we need to know that a praised approach is used by the authors, information which may not be local in the paragraph).

Appendix A gives unambiguous example cases where the citation function can be decided on the basis of the sentence alone, but Fig. 4 shows a more typical example where more context is required to interpret the function. The evaluation of the citation *Hindle (1990)* is contrastive; the evaluative statement is found 4 sentences after the sentence containing the citation[3]. It consists of a positive statement (agreement with authors' view), followed by a weakness, underlined, which is the chosen category. This is marked on the nearest markable (*Hindle*, 3 sentences after the citation).

[2] In all corpus examples, numbers in brackets correspond to the official Cmp_lg archive number, "S-" numbers to sentence numbers according to our preprocessing.

[3] In Fig. 4, markables are shown in boxes, evaluative statements underlined, and referents in bold face.

S-5 **Hindle (1990)**/Neut *proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen.*
S-6 *For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs.*
S-7 *This requires a reasonable definition of verb similarity and a similarity estimation method.*
S-8 *In* **Hindle**/Weak *'s proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.*
S-9 **His** *notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.* (9408011)

Figure 4: Annotation example: influence of context

A naive view on this annotation scheme could consider the first two sets of categories in our scheme as "negative" and the third set of categories "positive". There is indeed a sentiment aspect to the interpretation of citations, due to the fact that authors need to make a point in their paper and thus have a stance towards their citations. But this is not the whole story: many of our "positive" categories are more concerned with different ways in which the cited work is useful to the current work (which aspect of it is used, e.g., just a definition or the entire solution?), and many of the contrastive statements have no negative connotation at all and simply state a (value-free) difference between approaches. However, if one looks at the distribution of positive and negative adjectives around citations, one notices a (non-trivial) connection between our task and sentiment classification.

There are written guidelines of 25 pages, which instruct the annotators to only assign one category per citation, and to skim-read the paper before annotation. The guidelines provide a decision tree and give decision aids in systematically ambiguous cases, but subjective judgement of the annotators is nevertheless necessary to assign a single tag in an unseen context. We implemented an annotation tool based on XML/XSLT technology, which allows us to use any web browser to interactively assign one of the 12 tags (presented as a pull-down list) to each citation.

3 Data

The data we used came from the CmpLg (Computation and Language archive; 320 conference articles in computational linguistics). The articles are in XML format. Headlines, titles, authors and reference list items are automatically marked up with the corresponding tags. Reference lists are parsed, and cited authors' names are identified. Our citation parser then applies regular patterns and finds citations and other occurrences of the names of cited authors (without a date) in running text and marks them up. Self-citations are detected by overlap of citing and cited authors. The citation processor developped in our group (Ritchie et al., 2006) achieves high accuracy for this task (96% of citations recognized, provided the reference list was error-free). On average, our papers contain 26.8 citation instances in running text[4].

4 Human Annotation: results

In order to machine learn citation function, we are in the process of creating a corpus of scientific articles with human annotated citations, according to the scheme discussed before. Here we report preliminary results with that scheme, with three annotators who are developers of the scheme.

In our experiment, the annotators independently annotated 26 conference articles with this scheme, on the basis of guidelines which were frozen once annotation started[5]. The data used for the experiment contained a total of 120,000 running words and 548 citations.

The relative frequency of each category observed in the annotation is listed in Fig. 5. As expected, the distribution is very skewed, with more than 60% of the citations of category Neut.[6] What is interesting is the relatively high frequency of usage categories (PUse, PModi, PBas) with a total of 18.9%. There is a relatively low frequency of clearly negative citations (Weak, CoCoR-, total of 4.1%), whereas the neutral–contrastive categories (CoCoGM, CoCoR0, CoCoXY) are slightly more frequent at 7.6%. This is in concordance with earlier annotation experiments (Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977).

We reached an inter-annotator agreement of K=.72 (n=12;N=548;k=3)[7]. This is comparable to aggreement on other discourse annotation tasks such as dialogue act parsing and Argumentative Zoning (Teufel et al., 1999). We consider the agreement quite good, considering the number of categories and the difficulties (e.g., non-local dependencies) of the task.

The annotators are obviously still disagreeing on some categories. We were wondering to what degree the fine granularity of the scheme is a problem. When we collapsed the obvious similar categories (all P categories into one category, and all CoCo categories into another) to give four top level categories (Weak, Positive, Contrast, Neutral), this only raised kappa to 0.76. This

[4] As opposed to reference list items, which are fewer.

[5] The development of the scheme was done with 40+ different articles.

[6] Spiegel-Rüsing found that out of 2309 citations she examined, 80% substantiated statements.

[7] Following Carletta (1996), we measure agreement in Kappa, which follows the formula $K = \frac{P(A)-P(E)}{1-P(E)}$ where P(A) is observed, and P(E) expected agreement. Kappa ranges between -1 and 1. K=0 means agreement is only as expected by chance. Generally, Kappas of 0.8 are considered stable, and Kappas of .69 as marginally stable, according to the strictest scheme applied in the field.

Neut	PUse	CoCoGM	PSim	Weak	CoCoXY	PMot	PModi	PBas	PSup	CoCo-	CoCoR0
62.7%	15.8%	3.9%	3.8%	3.1%	2.9%	2.2%	1.6%	1.5%	1.1%	1.0%	0.8%

Figure 5: Distribution of the categories

	Weak	CoCo-	CoCoGM	CoCoR0	CoCoXY	PUse	PBas	PModi	PMot	PSim	PSup	Neut
Weak	**5**											3
CoCo-		**1**	3									
CoCoGM			**23**							3		
CoCoR0				**4**								
CoCoXY					**1**							
PUse						**86**	6			2	1	12
PBas							**3**					2
PModi								**3**				
PMot									**13**			4
PSim						3				**20**		5
PSup		1				2					**1**	
Neut	6		10	6	4	17	1		6	4		**287**

Figure 6: Confusion matrix between two annotators

points to the fact that most of our annotators disagreed about whether to assign a more informative category or `Neut`, the neutral fall-back category. Unfortunately, Kappa is only partially sensitive to such specialised disagreements. While it will reward agreement with infrequent categories more than agreement with frequent categories, it nevertheless does not allow us to weight disagreements we care less about (`Neut` vs more informative category) less than disagreements we do care a lot about (informative categories which are mutually exclusive, such as `Weak` and `PSim`).

Fig. 6 shows a confusion matrix between the two annotators who agreed most with each other. This again points to the fact that a large proportion of the confusion involves an informative category and `Neut`. The issue with `Neut` and `Weak` is a point at hand: authors seem to often (deliberately or not) mask their intended citation function with seemingly neutral statements. Many statements of weakness of other approaches were stated in such caged terms that our annotators disagreed about whether the signals given were "explicit" enough.

While our focus is not sentiment analysis, it is possible to conflate our 12 categories into three: *positive*, *weakness* and *neutral* by the following mapping:

Old Categories	New Category
Weak, CoCo-	Negative
PMot, PUse, PBas, PModi, PSim, PSup	Positive
CoCoGM, CoCoR0, CoCoXY, Neut	Neutral

Thus negative contrasts and weaknesses are grouped into `Negative`, while neutral contrasts are grouped into `Neutral`. All the positive classes are conflated into `Positive`. This resulted in kappa=0.75 for three annotators.

Fig. 7 shows the confusion matrix between two annotators for this sentiment classification. Fig. 7 is particularly instructive, because it shows that annotators

	Weakness	Positive	Neutral
Weakness	**9**	1	12
Positive		**140**	13
Neutral	4	30	**339**

Figure 7: Confusion matrix between two annotators; categories collapsed to reflect sentiment

have only one case of confusion between positive and negative references to cited work. The vast majority of disagreements reflects genuine ambiguity as to whether the authors were trying to stay neutral or express a sentiment.

Distinction	**Kappa**
`PMot` v. all others	.790
`CoCoGM` v. all others	.765
`PUse` v. all others	.761
`CoCoR0` v. all others	.746
`Neut` v. all others	.742
`PSim` v. all others	.649
`PModi` v. all others	.553
`CoCoXY` v. all others	.553
`Weak` v. all others	.522
`CoCo-` v. all others	.462
`PBas` v. all others	.414
`PSup` v. all others	.268

Figure 8: Distinctiveness of categories

In an attempt to determine how well each category was defined, we created artificial splits of the data into binary distinctions: each category versus a super-category consisting of all the other collapsed categories. The kappas measured on these datasets are given in Fig. 8. The higher they are, the better the annotators could distinguish the given category from all the other categories. We can see that out of the informa-

tive categories, four are defined at least as well as the overall distinction (i.e. above the line in Fig. 8: `PMot`, `PUse`, `CoCoGM` and `CoCoR0`. This is encouraging, as the application of citation maps is almost entirely centered around usage and contrast. However, the semantics of some categories are less well-understood by our annotators: in particular `PSup` (where the difficulty lies in what an annotator understands as "mutual support" of two theories), and (unfortunately) `PBas`. The problem with `PBas` is that its distinction from `PUse` is based on subjective judgement of whether the authors *use* a part of somebody's previous work, or *base* themselves entirely on this previous work (i.e., see themselves as following in the same intellectual framework). Another problem concerns the low distinctivity for the clearly negative categories `CoCo-` and `Weak`. This is in line with MacRoberts and MacRoberts' hypothesis that criticism is often hedged and not clearly lexically signalled, which makes it more difficult to reliably annotate such citations.

5 Conclusion

We have described a new task: human annotation of citation function, a phenomenon which we believe to be closely related to the overall discourse structure of scientific articles. Our annotation scheme concentrates on contrast, weaknesses of other work, similarities between work and usage of other work. One of its principles is the fact that relations are only to be marked if they are explicitly signalled. Here, we report positive results in terms of interannotator agreement.

Future work on the annotation scheme will concentrate on improving guidelines for currently suboptimal categories, and on measuring intra-annotator agreement and inter-annotator agreement with naive annotators. We are also currently investigating how well our scheme will work on text from a different discipline, namely chemistry. Work on applying machine learning techniques for automatic citation classification is currently underway (Teufel et al., 2006); the agreement of one annotator and the system is currently K=.57, leaving plenty of room for improvement in comparison with the human annotation results presented here.

6 Acknowledgements

This work was funded by the EPSRC projects CITRAZ (GR/S27832/01, "Rhetorical Citation Maps and Domain-independent Argumentative Zoning") and SCIBORG (EP/C010035/1, "Extracting the Science from Scientific Publications").

References

Susan Bonzi. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *JASIS*, 33(4):208–216.

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

Daryl E. Chubin and S. D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):423–441.

Carolyn O. Frost. 1979. The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly*, 49:405.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, pages 89–98.

Jorge E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United Stated of America (PNAS)*, 102(46).

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

Michael H. MacRoberts and Barbara R. MacRoberts. 1984. The negational reference: Or the art of dissembling. *Social Studies of Science*, 14:91–94.

Michael J. Moravcsik and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5:88–91.

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

John O'Connor. 1982. Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, 18(3):125–131.

Charles Oppenheim and Susan P. Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *JASIS*, 29:226–230.

Anna Ritchie, Simone Teufel, and Steven Robertson. 2006. Creating a test collection for citation-based IR experiments. In *Proceedings of HLT-06*.

Simon Buckingham Shum. 1998. Evolving the web for scientific knowledge: First steps towards an "HCI knowledge web". *Interfaces, British HCI Group Magazine*, 39:16–21.

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

John Swales. 1986. Citation analysis and discourse analysis. *Applied Linguistics*, 7(1):39–56.

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

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

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, UK.

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

Howard D. White. 2004. Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25(1):89–116.

John M. Ziman. 1968. *Public Knowledge: An Essay Concerning the Social Dimensions of Science*. Cambridge University Press, Cambridge, UK.

A Annotation examples

Weak	*However,* **Koskenniemi** *himself understood that his initial implementation had significant limitations in handling non-concatenative morphotactic processes.* (0006044, S-4)
CoCoGM	*The goals of the two papers are slightly different:* **Moore** *'s approach is designed to reduce the total grammar size (i.e., the sum of the lengths of the productions), while our approach minimizes the number of productions.* (0008021, S-22)
CoCoR0	*This is similar to results in the literature* **(Ramshaw and Marcus 1995).** (0008022, S-147)
CoCo-	*For the Penn Treebank,* **Ratnaparkhi (1996)** *reports an accuracy of 96.6% using the Maximum Entropy approach, our much simpler and therefore faster HMM approach delivers 96.7%.* (0003055, S-156)
CoCoXY	*Unlike previous approaches* **(Ellison 1994, Walther 1996), Karttunen** *'s approach is encoded entirely in the finite state calculus, with no extra-logical procedures for counting constraint violations.* (0006038, S-5)
PBas	*Our starting point is the work described in* **Ferro et al. (1999)** *, which used a fairly small training set.* (0008004, S-11)
PUse	*In our application, we tried out the Learning Vector Quantization (LVQ)* **(Kohonen et al. 1996).** (0003060, S-105)
PModi	*In our experiments, we have used a conjugate-gradient optimization program adapted from the one presented in* **Press et al.** (0008028, S-72)
PMot	*It has also been shown that the combined accuracy of an ensemble of multiple classifiers is often significantly greater than that of any of the individual classifiers that make up the ensemble (e.g.,* **Dietterich (1997)).** (0005006, S-9)
PSim	*Our system is closely related to those proposed in* **Resnik (1997)** *and* **Abney and Light (1999).** (0008020, S-24)
PSup	*In all experiments the SVM_Light system outperformed other learning algorithms, which confirms* **Yang and Liu** *'s* **(1999)** *results for SVMs fed with Reuters data.* (0003060, S-141)
Neut	*The cosine metric and Jaccard's coefficient are commonly used in information retrieval as measures of association* **(Salton and McGill 1983).** (0001012, S-29)