Models of Scholarly Communication and Citation Analysis

Fredrik Åström[1] and Ágnes Sándor[2]

[1] *fredrik.astrom@lub.lu.se*
Lund University Libraries, P.O. Box 134, SE-221 00 Lund (Sweden)
University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007 (Australia)

[2] *Agnes.Sandor@xrce.xerox.com*
Xerox Research Centre Europe, 6. chemin Maupertuis, F-38240 Meylan (France)

Abstract
Informetric/bibliometric analyses have to a large extent been relying on an assumption that research is essentially cumulative in its nature, which is not the least visible in the rational for using citation analyses to assess quality of research. However, when reviewing both the theoretical literature on how research is organized and studies analyzing the structures of research fields through informetric mapping methods, it becomes clear that cumulative organization is just one category of several ways of organizing research and scholarly communication, Consequently, the way the role of citations is interpreted in research assessment has to be revised. Based on the review of previous research, this paper suggests a model for categorizing different modes of scholarly communication. We test this model through three different kinds of semantic labelling analyses on abstracts and research papers from the fields of biomedicine, computer science and educational research. The model proposed suggests three main categories of scholarly communication: cumulative, negotiating and distinctive; and when matching the labels identified in the semantic analysis to the three categories, we find evidence of the three different ways of communicating research that supports the model.

Introduction
Citation analysis has become an increasingly important and utilized tool for evaluating research quality through the development and use of various impact indicators for e.g. journals and universities. And while many of the problems associated with the use of these indicators – such as variations between research fields in terms of coverage in databases and publication practices, as well as differences between publication types and the impact of 'negative' citations – has been addressed, the underlying assumptions for citation analysis has been discussed to a lesser extent. The basic assumption for relating citations with quality is that research is essentially cumulative: the texts cited are the foundation on which new research is built and are chosen because of their high quality. Furthermore, it is also assumed that research is to a large extent organized on a subject or disciplinary basis, i.e. researchers build on previous research conducted within their own fields, and to a large extent on relatively contemporary work. However, this is not the only way of organizing research and scholarly communication, as we can see both when looking at the theoretical literature on research organization and informetric mapping of research fields.

As with the underlying assumption of a cumulatively organized research as basis for the citation analysis rational, the relation between citation analysis and informetrics/bibliometrics on one hand, and different modes of research organization and scholarly communication on the other, has not been studied to any greater extent. While there have been many studies on the different types of citations and citing motivations, they have not tried to relate the various kinds of citations with different forms of scholarly communication.

Thus, the purpose of this paper is to suggest a model for categorizing different forms of scholarly communication based on theories on how research is organized, as well as on the results of mapping analyses of different research fields, and to test the model through a set of semantic analyses of relations between texts from three different research fields to investigate if we can find evidence of different modes of scholarly communication within the texts.

Proceedings of ISSI 2009, pages 10–21.
Edited by B. Larsen and J. Leta.

Modelling scholarly communication

Science Studies

In 1942, Robert K Merton described the 'ethos of science', captured in the acronym CUDOS for 'communism', 'universalism', 'disinterestedness' and 'organized scepticism', a complex of norms and values directing how research should be conducted (Merton, 1974). Few might argue against the principles expressed in this ethos, although many scholars from the 1950s and onwards would claim them unattainable, due to problems of setting one complex of regulating norms for sets of activities that are widely varying in terms of intellectual and social organization. And while Merton, together with Thomas Kuhn, are 'founding fathers' of the study of the social aspects of science, both have contributed to solidifying a strong normative view on the organization of scientific enterprises: Merton through the formulation of the ethos and Kuhn by distinguishing between mature and immature fields of research (Kuhn, 1970). This tendency towards implementing a strong norm for how research should be organized and conducted has been going on throughout the 20th century, and is largely based on an idealized model of the hard sciences characterized e.g. by the cumulative nature and a focus on the use of quantitative methods.

The norm of cumulatively organized and quantitatively performed sciences has become hugely influential in shaping how research is regarded – in general as well as within the scientific community – and not the least, in the use of citation analysis to measure impact or quality of research. Whether impact is indicated by the Journal Impact Factor, the *h*-index or normalized field citation scores, the underlying assumption is that research is essentially cumulative, and citations are essentially included to refer to the literature that the research presented is built upon. Merton discussed this by linking the normative nature of science with the reward system and how e.g. excellence is recognized and science is evaluated (Merton, 1974). Many problems associated with this assumption have been identified: Merton discussed the skewed distribution of recognition when writing about the 'Matthew effect' (Merton, 1968); the variations in types of citations as well as citing motivations are well known (e.g. Brooks, 1985; Moravcsik & Murugesan, 1975) as well as variations in citation rates in different research fields, publication types and so on (McRoberts & McRoberts, 1989); and the literature on the role, nature and meaning of citations is substantial (e.g. Cronin, 1984; Leydesdorff, 1998; Small, 1978; Wouters, 1999).

Meanwhile, in the social studies of science, the normative view on scientific research has been challenged by addressing differences between research fields (Becher & Trowler, 2001; Knorr Cetina, 1999, Whitley, 2000), non-disciplinary organized research (Klein, 1990) and the general development of science and research since 1945 (Gibbons et.al., 1994). One issue has been to avoid the distinction between social and intellectual aspects of scientific activities and instead show how they interact. Another issue has been to circumvent the 'Kuhnian' dichotomization between mature and immature fields of research and replace a normative view on research being based on *one* model of research organization with a model for analyzing research taking into account the various differences in the social and intellectual organization of various scientific fields.

One such model has been suggested by Richard Whitley (2000), where the social and intellectual dimensions of different research fields are analyzed through variations in terms of to what extent scholars – in order to get access to resources – need to show adherence to established methods, techniques and terminologies and demonstrate the coordination to the wider intellectual goals of the field. Based on Whitley's model, we can investigate how the organization of scientific activities differs between research fields, e.g. in terms of to what extent the results are predictable, if there are collective standards on what methods, theories and terminology to use, as well as if there are agreements on what research problems that are

worthy of being pursued, and whether – and in what way – the researchers need to persuade colleagues about the importance of their research. Based on this model, we can study e.g. to what extent results – and the evaluation of results as well as research proposals – are subject to different interpretations, and – of vital importance for how to interpret the results of citation analyses – to what extent and in what way we relate to our colleagues within the field when addressing a research problem.

Visualization studies

The assumption that researchers cite texts by peers within their own fields in a cumulative manner is not only a precept for citation analysis for research evaluation purposes, but is a basic assumption for informetric/bibliometric analyses in general, including co-citation analyses (Marshakova, 1973; Small, 1973) and author co-citation analysis (ACA) (McCain, 1986; White & Griffith, 1981) for the purpose of mapping intellectual structures of research fields. By building onwards on the literature specific to the researchers' own specialties, links between documents and authors through citations amount to representing an intellectual affinity, which means that by studying these links on a quantitative basis, we can identify different specialties within a research field. However, when applying these methods on different fields, the variations in what kind of results we get and how they can be interpreted, are quite substantial. This can be seen as reflections of different ways of organizing research and scholarly communication.

One aspect of the variation between fields is the consistency of the results, i.e. to what extent variations in terms of methods used and material being analyzed yields the same structure of a particular field. When studying a field like genetics (Åström, 2008b), the results form clearly identifiable structures and are quite homogeneous, independent of which channels of communication are analyzed. The same is not the case when analyzing fields like library and information science (LIS) and XML research, where the structures differ depending on e.g. what material is analyzed (Åström, 2002; White & McCain, 1998; Zhao, 2004). In the case of LIS we also find differences depending on levels of analysis (Moya-Anegón et.al., 2006). And while an ACA of comparative literature research (CLR) yields some – although not very clear – results (Hellqvist, 2008), an analysis of co-citations restricted to citations among contemporary CLR scholars as suggested by Persson (2001) shows almost no results (Hellqvist & Åström, 2007). A similar case can also be found when studying educational research (Åström, 2008a), where the ACA map primarily features citation classics, while the restricted analysis only features a few narrow specialities.

To some extent, these traits can be explained by different publication practices: since much of the results of CLR research is published in books, a restricted analysis of CLR journal articles indexed in the ISI databases would not include the research presented in other media. However, when studying the results of the unrestricted analyses, we find CLR citation classics to some extent, but the majority of cited authors are general theorists from fields like philosophy, sociology and linguistics whereas contemporary CLR scholars are almost non-existent (Hellqvist, 2008).

Model

In social studies of science, Whitley (2000) developed a model for analyzing the intellectual and social structure of research fields. An important part of that model is the concept of 'mutual dependency', reflecting the social aspects of research organization through variations in to what extent scholars are depending on their peers to make significant contributions to the intellectual goals of the research field. This involves aspects such as to what extent familiar and commonly accepted techniques, methods and materials needs to de utilized, as well as the level of consensus on competence standards and assessment criteria; and to what extent these

issues needs to be addressed when persuading peers of the importance of their research problem. Thus, if the level of mutual dependency is high, the scholars needs to address their connection and level of contribution to the research field by showing how they conform to accepted work practices within the field; whereas in a field with a low level of mutual dependency, issues like coordination of work practices does not need to be addressed in the same way and the contribution to the research field might to a much larger extent be motivated by demonstrating the uniqueness of the research in terms of e.g. methods used or material studied. In terms of citation analysis, this has implications in terms of to what level scholars need and will be citing their colleagues. Based on Whitley's (2000) model for understanding the organization of research and the variations in terms of how homogeneous different fields are in terms of methods, theories and terminologies, as well as to what extent scholars needs to demonstrate their adherence to techniques and relation to colleagues; and the varying results of the co-citation analyses on different research fields and how these results can be interpreted, we can identify three main categories of scholarly communication (Figure 1.). It should be noted that these categories are not mutually exclusive: different research fields, as well as individual texts, show traits of all three categories. However, based on the theoretical literature as well as on the results of different mapping studies, we do assume that individual fields are primarily related to one particular category.

Cumulative research, the first category, is of course the type of organization of research and scholarly communication that has been the norm for research since more than a century, where the relation between texts is marked by a continuous progression within the field: building on previous literature to make new discoveries. As opposed to the cumulative processes identified in the first category, much research is also primarily performed and presented by *negotiating* the interpretation of different phenomena. Whereas the first two categories are characterized by relations between texts, there is also a category where the main function is to show the *distinctive* trait in a line of research: either by in some way 'disqualifying' previous research or by showing the originality of the present research. This category can be expressed both by e.g. 'negative' citations or, when it comes to the category as a trait in the characterization of a research field, by not really relating to other literature within the field because the research is assumed to be so unique to render stated relations to peers superfluous.

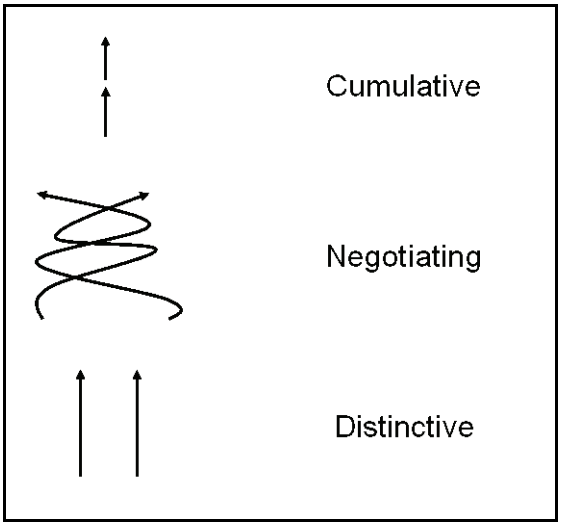

Figure 1. Model of scholarly communication categories

Within these categories, we find distinct variations in terms of what citations mean and how citation analyses can be interpreted, i.e. to what extent citations can be seen as an indicator of quality and how interpret a 'negative' citation can be interpreted. In research being characterized by a cumulative organization and communication, most citations can be seen as an indicator of quality, since we're assuming that when a researcher chooses which research to build onwards on, (s)he chooses good research. However, if (s)he cites something in a negative way, it is because the research is in some way faulty or erroneous, and thus of low quality. If the communication within a field is primarily of a negotiating kind, the link between citations and quality is not necessarily as strong as in the first case. A researcher might cite literature that is hugely influential and of high quality, but suggest alternative interpretations. Such a negative citation does not necessarily mean that the cited text is of low quality. On the contrary it signals that the cited text is relevant as opposed to a negative citation signalling errors in results or method. In fields where communication is primarily distinctive rather than relational, we lose even more of the connection between citations and quality indicators, since the citation traffic is essentially going to literature outside the researcher's own field, and the impact of the research is not reflected in citations by peers but in other ways.

As mentioned before, these different ways of communication within different fields do not only reflect varying citation practices and relate to how we can understand citations in texts from different fields, but they are also related to the wider organization of the field, to how research is evaluated and to how researchers justify and motivate their contributions to the intellectual goals of the field. In a cumulative setting, the role of citations is partly to justify new contribution to the field by showing adherence to generally acknowledged standards for research techniques and methods through building on earlier research efforts to discover new phenomena. In a field characterized by communication largely being of a negotiating kind, contribution to the field is still justified by relating to other researchers and their texts, but at the same time also by discussing and suggesting alternate interpretations of known phenomena. In research where communication is largely of a distinctive nature, there is no need to relate to what other scholars in the field are doing since the researcher is largely pursuing his(her) own research agenda and his(her) contribution to the field is largely motivated by the uniqueness of the research effort in terms of e.g. empirical material or theoretical approach.

The model is of course an idealization, providing an interpretational structure for analyzing different modes of scholarly communication and research organization, where traits from the different ways of communication can be found in most fields of research. And it should be noted that the semantic analyses being performed in this analysis does not aim towards classifying research fields as being primarily operating inside one particular mode of communication; the aim is rather to find evidence of these different forms of communication.

Method

The construction of the model of scholarly communication categories is based on the one hand on an analytical review of research on the intellectual and social organization of science and research, and on the other on studies mapping and visualizing the intellectual structure of different fields of research through informetric methods.

To test the model: corpora of texts from the fields of educational research, biomedicine and computer science were analyzed, using the natural language processing tool Xerox Incremental Parser (XIP) to identify evidence of different relations between previous and current texts within the three fields. Three different kinds of semantic labelling of the text corpora were carried out: the first identifies evidence of breaks with previous research in biomedicine article abstracts, the second identifies relationships between cited and citing

articles in computer science by analyzing the sentences in which the citations occur and the third identifies citation types in educational research literature by an analysis similar to the one on computer science. After identifying evidence of these various relationships between texts in the fields, the different kinds of relationships are matched to the three categories of scholarly communication.

Testing the model

We test the model independently of the methods which have been applied to build it, and which are described above. Our test relies on recent developments of automatic semantic processing (Sándor, 2006) that enable us to provide empirical evidence of the existence of the categories by analyzing the embodiment of scholarly communication, i.e. the texts of publications. In previous work fine-grained automatic semantic analysis of scientific discourse has been used for detecting various aspects of the content of scientific articles such as: information retrieval and extraction (e.g. the ACE project: (http://projects.ldc.upenn.edu/ace/), rhetorical analysis (Mizuta et.al., 2006), summarization (Teufel et.al., 2002) and citation analysis (Teufel et.al., 2006), but to our knowledge, ours is the first attempt to apply it for the detection of the categories of scholarly communication.

Three different text analysis systems have recently been developed at the Xerox Research Centre Europe as parts of past and ongoing research projects for the semantic tagging of scientific articles. Although the three systems have been developed for applications that are different both from each other and from the goals of the present article (the original applications being text mining, semantic mark-up and research quality evaluation), we can take advantage of the output of the semantic tagging they provide since they can be linked to the categories of our model of scholarly communication in a relatively straightforward way.

The three applications detect and label sentences that express some kind of relationship between previous work and the current work. The types of relationships detected in the three applications are different from each other. However, since they are all conceptually lower-level relationships than the ones in our model, they can be mapped in one of those relationships.

Our long-term goal is to build an analysis tool whose coverage and precision is high enough to provide statistical data concerning the proportion of the different modes of communication across scientific disciplines. The results presented here do not label all the sentences that convey one of the three categories in the model. This is due to the fact that the three systems have been designed for detecting conceptually lower-level relationships. Thus in this paper, by using some existing systems we restrict ourselves to claiming to show that current semantic analysis systems are fine-grained enough for providing evidence of the existence of the categories, and they also provide some preliminary statistical data. However, since they do not cover the three categories proposed, we cannot use them at this point for overall large-scale testing.

All the three systems have been developed applying syntactic (Hagège et.al., 2002) and semantic (Brun et.al., 2003, 2004) parsing as well as the concept-matching framework (Sándor, 2007) developed at the Xerox Research Centre Europe. They all rely on models of linguistic expressions conveying the descriptions of relationships among scientific articles. The models are encoded in the Xerox Incremental Parser (Aït et.al., 2001; 2002). The parser carries out dependency analysis of sentences and some general semantic analysis (named entity extraction, basic semantic normalization). It is equipped with a rule writing formalism that enables grammar-writers to enrich existing analyses with further, more fine-grained analyses.

In the following paragraphs we describe the three systems and map the results into the three high-level categories of the above-mentioned model of scholarly communication: cumulative,

negotiating and distinctive. The three systems have been worked out for three different research fields: biomedicine, computer science and educational sciences.

Biomedicine

The semantic labelling system in the field of biomedicine was developed in the framework of a joint research project in the field of text-mining (Lisacek et.al., 2005). The purpose of the project was to detect biomedical research abstracts in Medline that describe research that breaks with traditional approaches, views, methods, etc. Abstracts usually do not contain citations, but the ones that we detected inherently break with previous research. Here are some examples of sentences detected in this application (the relevant expressions are underlined):

> *In contrast with previous hypotheses, compact plaques form before significant deposition of diffuse A beta, suggesting that different mechanisms are involved in the deposition of diffuse amyloid and the aggregation into plaques.*
>
> *Models of neurodegenerative disorders are challenging the classical defining role of tangles in neurotoxicity.*
>
> *The patterns of neurotoxicity possessed a complex pharmacological profile, demonstrated an apoptotic-necrotic continuum and were inconsistent with past findings.*

In the application an abstract is labelled if it contains expressions describing substantially new results. The system was evaluated to be fairly precise linguistically: 92% of the abstracts detected do in reality contain expressions that indicate that the results presented are substantially new. We do not have any measures for its coverage. (For the details of the evaluation see Lisacek et.al., 2005)

As the examples above show the labelled abstracts do not represent a cumulative communication model. They could be either negotiating or distinctive with respect to our system of scholarly communication.

We tested the filtering power of the system on a corpus of 3300 abstracts retrieved with some subject matter keywords. The system detected 175 abstracts, which is 5% of the original collection. Thus according to this test at least 5% of biomedical research abstracts describe research that negotiates or is distinctive from previous work.

Computer science

The system that we are going to briefly describe here was developed within the Sixth Framework Integrated European Project (Vikef [http://www.vikef.net/), where one of the aims was the semantic enrichment of scientific articles. The special task relevant for the model of scholarly communication is labelling the sentences in which previous work is cited according to the relationship between the citing and the cited paper. We distinguished 4 types of relationships:

- BACKGROUND_KNOWLEDGE: the cited work provides background knowledge to understanding the citing work
- BASED_ON: the citing work is based on the cited work
- COMPARISON: comparison between the citing and the cited work
- ASSESSMENT: assessment of the cited work

We can map the semantic labels of citations to the categories of scholarly communication in the following way:

- BASED_ON is surely cumulative
- COMPARISON is cumulative or negotiating
- ASSESSMENT: negative assessment is negotiating

We cannot say anything about BACKGROUND_KNOWLEDGE and ASSESSMENT when it is positive. Again, we illustrate the relationship types with some examples (we do not have examples of negative assessment in our corpus):

BASED_ON:

> *A key aspect of the approach we take is to apply the principle of data independence [13] by separating a logical from a physical layer.*
>
> *Although Protege also has some support for the Abstract Syntax , we chose to develop an expression syntax based on standard DL symbols [1].*
>
> *A previous paper on modeling messages in RDF [3] proposed an ontology for messaging, and we have reused the portions of this ontology that are applicable to blogging.*

COMPARISON:

> *Lemma 3.2 below shows that exactly the same class of rdfs-interpretations is defined as in [2].*
>
> *We had also considered TRIPLE [24] and Racer.*
>
> *Indeed, some of the constructs presented there, drawn from earlier work by Guha [10] and McCarthy [15], are very similar in spirit to those presented here.*

This system also provides some evidence of the categories, but we do not have overall statistics of the representativity of the categories in the field. 7% of the citations was labelled as BASED_ON, which means that at least 7% is cumulative, and 2,5% was labelled as COMPARISON, which means that at least 2,5% of the citations are cumulative or negotiating.
The system was evaluated on a corpus of 511 citation sentences: the recall was measured to be 86%, i.e. 86% of the citations were labelled correctly compared with human annotation, and the precision was 92%, i.e. 92% of the labels assigned by the system were right according to the human annotation.

Educational sciences

The system for assigning semantic labels to articles in the field of educational sciences is being constructed in the ongoing EERQI European project within 7th Framework Programme for Research in the Socio-economic Sciences and Humanities Theme (http://www.eerqi.eu/). The aim of the project is developing quality indicators for educational sciences. Semantic labelling of citation types is experimented as a semantic indicator of quality of the cited work. Similarly to the Vikef project we are setting up citation types and constructing the models of

sentences that convey them. The citation types that are being determined are different from those in computer science. Their list at present is the following:

- ARGUMENTATION: argumentation between the citing and the cited work
- EVIDENCE: the cited work provides evidence for the cited work
- IMPORTANCE: the author of the citing work finds the cited work important
- QUALIFICATION: the cited work is qualified by the citing work
- SURPRISE: the author of the citing work is surprised by the cited work

Out of these labels ARGUMENTATION could be linked to the negotiating category of scholarly communication and EVIDENCE to the cumulative category. The following examples illustrate the type of sentences relevant for our categories:

ARGUMENTATION:

> *I aim in this paper to shift focus and suggest that rather than seeing the world of young people as defined by the wicked ' faceless people in the virtual world of cyberspace ' (Ziegler 2007 , 76) , the ' Pied Piper ' who will seduce your children (78) , or the ' wolf in sheep 's clothing ' (69) that will sneak into your home , we might do better to recognise that these cybersites can be productive and agentic spaces within which young people , especially girls , might be engaged in their own important work .*
>
> *The many discussions between the different theoretical positions of constructivism , all with varying emphasis , have inhibited the narrowing of the bridge between theory and practice (Kennedy , 1997) .*
>
> *Outside of reading Bourdieu for method and to guide research , I am hesitant to use his concepts to theorise , which makes it unlikely that I should be writing about something called a ' rural habitus ' .*

EVIDENCE:

> *Evidence suggests that boys masculinities become shored up and exemplified in the most polarised and misogynistic ways by boys who collectively feel undermined in and out of school and respond by taking on violent positions (Frosh , Phoenix and Pattman 2002) .*
>
> *To support this claim I can only quote some examples , both from 4 / 98 and from Hay McBer:*
>
> *As WOOD 's (2003) critical assessment of the U.S. diversity debate proves , since 1978 , different levels of court rulings establishing affirmative action and equal employment opportunity schemes in public institutions , organizations and businesses have forced both public and private actors to introduce diversity mechanisms into their particular organizational contexts .*

Since this work is ongoing we do not have any measures to report concerning the performance of the system or the proportion of the presence of the categories among the sentences that

contain citations. However, at this stage we do provide evidence of the non-cumulative use of previous work.

Conclusions

Science studies, visualization studies of the structure of research fields, as well as evaluations of science through citation analyses assume that research is basically cumulative. A close look at the theoretical literature indicates that this assumption is controversial, so we replace this assumption by a three-partite model of scientific communication, in which besides being cumulative, the organization of, and communication in, research can also be of a negotiating or distinctive nature. The results of automatic semantic analysis support our suggestion since we can give evidence of scientific communication being negotiating and distinctive.

Since we have not performed the same kind of analyses on the three fields we can say little about any overall differences between the fields or whether any of the fields can be primarily related to one specific type of communication. The semantic analysis does, however, show evidence of at least one kind of distinctive communication, where the author outspokenly distinguishes himself from previous results or research strategies. In contrast to this the kind of distinctively organized research suggested by the co-citation analyses of e.g. comparative literature could not be traced in our examples. One issue is of course the limitation of our experiments to three research fields, none of which belong to the arts and humanities, like e.g. comparative literature. The other issue in terms of identifying the distinction performed by not citing contemporary peers within the field is of course related to the problems of detecting something that is not there, i.e. the citations to the peers within their own field.

These problems of detecting what is not there is of course a limitation of semantic labelling as a method for identifying different kinds of scholarly communication. However, as demonstrated in the overview of variations in results of co-citation and visualization analyses, in this case the more traditional citation analysis can contribute by showing what is and what is not there, thus we can see how the two different methods can be combined to identify basic traits in how research is communicated. Furthermore, taking into account both the theories of Whitley and the differences between research fields and the results of previous results of research visualization studies, we would suggest that different kinds of research fields can be seen as primarily relating to one of the suggested modes of scholarly communication, including e.g. to what extent citations are made to build on previous research or to negotiate interpretations of phenomena, as well as to what extent citations are made within the field or to literature from other fields.

The implications of this in terms of citation analyses for research quality assessment are of course that they have to be used with great care and with a thorough understanding of how research and scholarly communication is organized in different fields of research. We suggest that these cultural and organizational differences among fields make it problematic to try to find one all encompassing model for evaluating research quality.

Acknowledgments

The research presented here was in part performed within European Educational Research Quality Indicators (EERQI). A project funded by the European Union's Seventh Framework Programme under the Socio-Economic Sciences and Humanities Theme (SSH).

References

Aït-Mokhtar, S., Chanod, J-P. & Roux, C. (2001). A multi-input dependency parser. In: *Proceedings of the Seventh IWPT (International Workshop on Parsing Technologies), Beijing, China, 17-19 October, 2001*. Beijing: Tsinghua University Press.

Aït-Mokhtar, S., Chanod, J-P. & Roux, C. (2002). Robustness beyond shallowness: Incremental dependency parsing. *Natural Language Engineering*, 8(2/3), 121-144.

Åström, F. (2002). Visualizing library and information science concept spaces through keyword and citation based maps and clusters. In: Bruce, Fidel, Ingwersen & Vakkari (Eds.) *Emerging frameworks and methods: CoLIS4* (pp. 185-197). Greenwood: Libraries Unlimited.

Åström, F. (2008a). Bibliometrics and educational research. Unpublished presentation: *The European Conference on Educational Research (ECER 2008) in Gothenburg, Sweden, 11 September 2008*.

Åström, F. (2008b). Citation patterns in open access journals. Unpublished project report: *The OpenAccess.se program, The National Library of Sweden*.

Becher, T. & Trowler, O.R. (2001). *Academic tribes and territories*. Buckingham: SRHE & Open University.

Brooks, T.A. (1985). Private acts and public objects. *The Journal of the American Society for Information Science*, 36(4), 223-229.

Brun, C. & Hagège, C. (2003). Normalization and paraphrasing using symbolic methods. In: *ACL: Second International workshop on Paraphrasing, Paraphrase Acquisition and Applications, Sapporo, Japan, July 7-12, 2003* (pp. 41-48).

Brun, C. & Hagège, C. (2004). Intertwining deep syntactic processing and named entity detection. *Advances in Natural Language Processing. Volume 3230* (pp. 195-206). Berlin/Heidelberg: Springer

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Gibbons, M. et.al. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London: Sage.

Hagège C. & Roux C. (2002). A robust and flexible platform for dependency extraction. In: *Proceedings of LREC 2002* (pp. 520-523). Las Palmas, Canaria, Spain.

Hellqvist, B. (2008). Bibliometri och humaniora: En relationsanalys [Bibliometry and the humanities: a relational analysis]. *InfoTrend*, 63(3).

Hellqvist, B. & Åström, F. (2007). Visualizing the humanities through co-citation analyses. Unpublished presentation: *12th Nordic Workshop on Bbibliometrics and Research Policy, Royal School of Library and Information Science, Copenhagen, Denmark, 13-14 September 2007*. Retrieved January 18, 2009 from: http://www.db.dk/nbw2007/files/5b_Hellqvist_%C5str%F6m.pdf

Klein, J.T, (1990). *Interdisciplinarity: History, theory and practice*. Detroit: Wayne State.

Knorr Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, Mass.: Harvard University.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1), 5-25.

Lisacek, F., Chichester, C., Kaplan, A. & Sándor, Á. (2005). Discovering paradigm shift patterns in biomedical abstracts: Application to neurodegenerative diseases. *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM)* (pp. 41-50). European Bioinformatics Institute, Hinxton, Cambridgeshire, UK, 10th -13th April, 2005.

MacRoberts, M.H. & MacRoberts, B.R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5), 342-349.

Marshakova, I. (1973). System of documentation connections based on references (SCI). *Nauchno-Tekhnicheskaya Informatsiya Seriya*, 2, 3-8.

McCain, K.W. (1986). Cocited author mapping as a valid representation of intellectual structure. *Journal of the American Society for Information Science*, 37(3), 111-122.

Merton, R.K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56-63.

Merton, R.K. (1974). *The sociology of science*. Chicago: The University of Chicago.

Mizuta, Y., Korhonen, A., Mullen, T & Collier, N. (2006), Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6), 468-487.

Moravcsik, M.J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86-92.

Moya-Anegón, F., Herrero-Solana, V., & Jiménez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: The SOM, clustering and MDS applied to library and information science research. *Journal of Information Science*, 32(1), 63-77.

Persson, O. (2001). All author co-citations versus first author co-citations. *Scientometrics*, 50, 339–344.

Sándor, Á., Kaplan, A. & Rondeau, G. (2006). Discourse and citation analysis with concept-matching. *International Symposium: Discourse and document (ISDD), Caen, France.* Retrieved 18 January, 2009 from: http://www.unicaen.fr/services/puc/article.php3?id_article=686

Sándor, Á. (2006). Using the author s comments for knowledge discovery. *Semaine de la connaissance, Atelier texte et connaissance, Nantes, June 29, 2006.* Retrieved 18 January, 2009 from: http://www.sdc2006.org/cdrom/contributions/Sandor.pdf

Sándor, Á. (2007). Modeling metadiscourse conveying the author s rhetorical strategy in biomedical research abstracts. *No. 2007-2 de la Revue Française de Linguistique Appliquée pp. 97-109.*

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327-340.

Teufel, S. & Moens, M. (2002) Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

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

White, H.D. & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-171

White, H.D. & McCain, K. (1998). Visualizing a discipline: an author cocitation analysis of information science. *Journal of the American Society for Information Science*, 49(4), 327-355

Whitley, R. (2000). The Intellectual and Social Organization of the Sciences. Oxford: University Press.

Wouters, P. (1999). *The citation culture*. (Doctoral Diss.), Universiteit van Amsterdam. Retrieved 18 January, 2009 from: http://garfield.library.upenn.edu/wouters/wouters.pdf

Zhao, D. (2004). *A comparative citation analysis study of web-based and print-journal based scholarly communication in the XML field*. (Doctoral Diss.), The Florida State University. Retrieved 18 January, 2009, from: http://etd.lib.fsu.edu/theses/available/etd-09232003-012028/