Citation Classification Based on Genre:
The Significance of the Textual Location of Citations

Master Thesis spring 2011
Royal School of Library and Information Science

Roald Erik Frøsig

Supervisor: Jeppe Nicolaisen

Co-citedness of research articles and the documents they cite, color-coded by location of the reference.
Shown on a double logarithmic scale.
The perfect line signifies citations = co-citations.
# Table of contents

1. **Introduction** ........................................................................................................... 6
   1.1 Hypotheses ............................................................................................................. 9
2. **Method** ..................................................................................................................... 10
   2.1 Source Articles ........................................................................................................ 10
   2.2 Determining the Rhetorical Moves ........................................................................... 13
   2.3 Comparing Documents ............................................................................................ 14
   Metadata ....................................................................................................................... 14
   Full Text ....................................................................................................................... 15
   Citation Analysis ......................................................................................................... 15
   2.4 Co-citations ............................................................................................................ 15
   Vector Space Model ..................................................................................................... 16
   2.5 Potential Methodical Problems ............................................................................... 17
   Age of References ........................................................................................................ 17
   Co-citation Analysis on Cited Documents .................................................................... 17
   Self Citations ............................................................................................................... 17
   Science Citation Index (SCI) ....................................................................................... 18
3. **Theory** ....................................................................................................................... 18
   3.1 Genre ..................................................................................................................... 18
   Genre Analysis ............................................................................................................. 19
   3.2 Discourse Communities ......................................................................................... 20
   3.3 The Scientific Community ...................................................................................... 20
   CUDOS ......................................................................................................................... 20
   Paradigms ..................................................................................................................... 21
   Social Constructivism ................................................................................................... 21
   3.4 Research Traditions and Specialties ...................................................................... 22
   3.5 The Research Article .............................................................................................. 23
   Purpose of the Research Article ................................................................................... 23
   Genre or Macrogenre? ................................................................................................. 24
   Subdivisions of the Research Article According to Swales ........................................ 24
   Dimensions Important for the Research Article ......................................................... 24
   Authorship ................................................................................................................... 25
   History of the Research Article .................................................................................... 26
Coded algorithmically ................................................................. 48
4.5 Discussion ................................................................................. 48
Moves .......................................................................................... 48
Suggestion .................................................................................. 49
Perspectives ............................................................................... 49
4.6 Conclusion .............................................................................. 50
References .................................................................................... 50
Appendices ................................................................................... 54
Appendix A: Bibliographic Information for the 30 Source Articles ....... 54
Appendix B: Move Structure Determined on Sentence Level ............... 56
Appendix C: Overview of Statistical Analysis ..................................... 81
References Grouped in Sections ..................................................... 81
Moves Coded by Sentences ............................................................. 84
Moves Coded by Paragraph ............................................................ 87
Moves Coded Algorithmically ......................................................... 90
Figures:
Fig. 1. The Hour-glass Model ................................................................. 8
Fig. 2. Illustration of co-citations and bibliographic couplings ...................... 16
Fig. 3. Distributions of citedness of references ........................................... 44
Fig. 4. Distribution of co-citedness between references and source articles ...... 45
Fig. 5. Co-citations compared with citation counts ...................................... 47
Fig. 6. Boxplot, sections ........................................................................... 81
Fig. 7. Histogram, sections ......................................................................... 83
Fig. 8. Boxplot, sentence-level moves ........................................................ 84
Fig. 9. Histogram, sentence-level moves ....................................................... 86
Fig. 10. Boxplot, paragraph-level moves ..................................................... 87
Fig. 11. Histogram, paragraph-level moves ................................................ 89
Fig. 12. Boxplot, algorithmic moves .......................................................... 90
Fig. 13. Histogram, algorithmic moves ........................................................ 92

Tables:
Table 1. Tripartitions for the algorithmic assignment of move. ....................... 14
Table 2. Characteristics of the 30 source articles. .......................................... 42
Table 3. Proposed additional metadata to enrich citation databases. ............... 49
Table 4. Correlations between Reference age and Vector based similarity ...... 81
Table 5. Descriptive data, sections .............................................................. 82
Table 6. Crosstabulation, sections ............................................................... 83
Table 7. Chi-square tests, sections ............................................................. 84
Table 8. Descriptive data, sentence-level moves .......................................... 85
Table 9. Crosstabulation, sentence-level moves .......................................... 86
Table 10. Chi-square test, sentence-level moves ......................................... 87
Table 11. Descriptive data, paragraph-level moves ...................................... 88
Table 12. Crosstabulation, paragraph-level moves ....................................... 89
Table 13. Chi-square test, paragraph-level moves ....................................... 90
Table 14. Descriptive data, algorithmic moves ............................................ 91
Table 15. Crosstabulation, algorithmic moves ............................................ 92
Table 16. Chi-square test, algorithmic moves ............................................. 93
Citation Classification Based on Genre: The Significance of the Textual Location of Citations

Roald Erik Frøsig
Royal School of Library and Information Science

Abstract. In this study, citation theory is combined with the genre concept in order to form a model for classification of citations from research articles. The classification system is based on the textual locations of the references. To test the applicability of classification by location, the references in 30 chemistry articles are grouped by section, and the CARS model is used to subdivide citations in the introductions. The results obtained using the Vector Space Model on the co-citation data gathered from the chemistry articles and their cited documents, show that locations of references can be used to predict similarity, that documents cited in introductions and method sections show markedly different co-citedness with the source article, and that recent documents cited at the end of research article introductions show a high degree of semantic relation to the citing article. Therefore the paper ends by suggesting that metadata on the textual location of references are added to citation databases.

Keywords: Citation Classification, Co-citation Analysis, Genre, Research Articles, Citation Theory, Metadata, CARS Model, Citation Indexes.

1 Introduction

An increasing amount of both practical and theoretical research has developed around citation analysis since the Institute for Scientific Information (ISI) established the Science Citation Index (SCI) in the 1960s, and made citation analyses on a grand scale feasible (MacRoberts 1989, Liu 1993, White 2004). The growing focus on evaluation of government-funded research has further increased the applications and consequences of citation analyses (Moed 2009).

Citation counts are frequently used as indicators of scientific impact or semantic relatedness of published research. The rationale behind using citations this way is founded on the following five assumptions (Smith 1981):

- Citation of a document implies use of that document by the citing author.
- Citation of a document reflects the merit of that document.
- Citations are made to the best possible works.
- A cited document is related in content to the citing document.
- All citations are equal.

The original purpose of SCI was to improve information retrieval (Garfield 2004), but a positive by-product was that it suddenly became feasible to make large-scale
citation analysis, which has proved itself to be an efficient tool for research evaluation (Cole & Cole 1973; Garfield 1994) and descriptive science studies (Small 1982).

However, in spite or perhaps because of the keen attention the field has received in the past few decades, a generally accepted theory of citation has failed to materialize, though several have been proposed (e.g. Gilbert 1977; Small 1978; Bavelas 1978; Wouters 1999).

Some have tried a more practical approach by introducing citation classification schemes (e.g. Moravcsik & Murugesan 1975; White & Wang 1997). The types of citations these schemes operate with have a certain degree of overlap (Small 1982), but the mutual agreement is limited to the fact that the significance of citations varies.

None of the theoretical or the practical propositions have won general acceptance, so we are left with a tool that works, though we are not sure why or how (Nicolaisen 2007).

One thing on which there is concurrence, is that Smith’s fifth assumption: All citations are equal, does not hold in reality (Cole & Cole 1971). The assumption refers to the practice in citation analyses to count the citations to a given document (or journal, or author, or institute, or nation), and let all citations weigh the same. This should be understood as a default practice forced by restrictions of reality, as no one can claim that all citations really are equal.

Evidently, it would open up for great improvements of citation analysis if it was possible to differentiate between different kinds of citations, especially if it could be done automatically (Swales 1986). It would then be possible to give different kinds of citations different weight allowing us to fine-tune algorithms based on citation counts. This would also open up for large-scale studies of the characteristics of citations and their relations to citing and cited documents, which would give us a better understanding of the nature of references and citations.

Automatic classification of citations have been tried before, current research on this centers on the words surrounding the reference in the text (e.g. Ritchie et al. 2008). As yet the results have not been convincing. An obstruction for this approach is that academic writing does not traditionally use overtly value-laden words, and opinions can be stated implicitly (Jacoby 1987; Swales 2004, p. 87), so that even readers with a scientific education but from another specialty, can misunderstand the meaning (MacRoberts & MacRoberts 1984; Swales 1986). Furthermore, a system built on this approach would be limited to one language.

Another procedure that might lead to computer-automated classification of citations is to count the number of times the cited documents are mentioned in the citing document (Voos & Dagaev 1976; Herlach 1978). There have also been studies that have looked at the location of the references in the citing documents, either in terms of section (introduction, method, results, discussion, conclusion), or relatively as a percentage of the total length of the citing document. The location approach was mentioned by Smith (1981), and has resurfaced a few times since. Peritz (1983) found that more than 50% of citations are located in the introduction, and that these primarily link to previous studies leading up to the current investigation. Cano (1989) also found that citations are concentrated in the beginning of papers, but that citations here tend to be perfunctory. Paul (2000) on the other hand, expected that citations in the introduction would be central to the article at hand, especially if they are repeated.
Apart from a few exceptions (e.g. Small 1978; Paul 2000) citation theorists as a rule do not focus on the location of the reference in the text, which is all the more striking as the reference has a rhetorical function at the point in the text where it is placed. It seems obvious to study citations in view of this function, and thus connect citation analysis with the concept of genre.

The most important genre in the world of science today is doubtless the research article, and the most citation-packed section of the research article is the introduction, therefore studying the citations of article introductions seems to be a promising starting place.

The characteristics of research articles have been thoroughly studied by linguists, the general picture is that the rhetoric has a broad perspective at the start of the introduction, and then narrows down towards the specific subject of the given article, following the hour-glass model as proposed by Hill et al. (Fig. 1).

**Fig. 1.** The Hour-glass Model. Illustrating the overall organization of the research article (Hill, Soppelsa & West 1982).

John. M. Swales (1981, 1990, and 2004) has designed a ‘create-a-research-space’ (CARS) model to describe the rhetorical structure of the introductions in research articles. This model consists of three rhetorical moves, which are subdivided in steps. The types and order of steps varies widely across different scientific fields, but the three moves are generally recognized as applicable in research articles in English across the fields (Swales 2004, p. 226). The three moves in the CARS model are:

1) *Establishing a territory* → 2) *Establishing a niche* → 3) *Occupying the niche*

In move 1, establishing a territory, the author draws attention to the article by accentuating the importance, potential and/or currency of the subject, usually followed by a review of existing knowledge, with many references. I expect that there will be numerous citations in this part of the introduction, that at least some of these
will be highly-cited documents suitable for getting the reader’s attention, and that the
documents cited here will be on the subject in general and not narrowly related to the
citing article.

Move 2 is to establish a niche; this can be done by indicating a gap in the current
knowledge, by questioning the validity of previous research, or by pointing out a
potential continuation of a current research tradition. This is often done by reporting
earlier research, and pointing out what is yet to be researched. Therefore I expect that
citations in move 2 will be to documents that are among the most closely related to
the citing article.

According to the hour-glass model the nearest relations of the current article
should be placed at the very end of the introduction, in move 3, but this move consists of
introducing the article and possibly describing the purpose, the structure and/or the
main findings of the research, so often there will not be any citations in this move
(Swales 1990, p. 159-161).

To my knowledge only one previous study has combined citation analysis with the
move-structure introduced by Swales. Paul (2000) found that two thirds of references
in introductions are located in move 1, and that these references tend to be asserted
stronger by the author than the references in move 2 and 3. The semantic relationship
between citing and cited document has not before been combined with the rhetorical
structure of the citing text.

Co-citations are regarded as a reliable measure of semantic relationship between
two documents (Small 1973; Egghe & Rousseau 1990, p. 242). Therefore I will use
co-citation analysis to evaluate the degree of relatedness between the cited and the
citing documents.

1.1 Hypotheses

The expressed hope therefore is that it should be possible to develop a system, or
more probably sets of systems, for categorizing citations so that their quality and
weight can be properly accounted for. If this could be done, so it is argued, citation
analysis would be a more reliable and a more equitable technique for carrying out
evaluative studies than it is at present (Swales 1986, p. 40).

A method that could classify citations meaningfully and automatically would
revolutionize citation analyses, both in the use of information retrieval, research
evaluation, and in the study of the sociology of science. Working from the assumption
that some citations points to documents that are more closely semantically related to
the citing article than others, I will try to answer the following question:

Is it possible to predict the degree of relatedness between a research article
and the documents it cites by the location of the reference in the citing article?

Based on the hour-glass model as described by Hill et al. (1982) and Swales (1990)
I will use the following general hypothesis:
Citations located in the end of the introduction will refer to documents that are most closely related to the citing article.

This study is designed after the create-a-research-space model, as it is described by Swales (1990, 2004). Therefore the precise hypotheses that I wish to test are the following:

- Move 1 will contain the highest number of references, but the documents cited here will have fewer co-citations with the citing article.
- Documents cited in move 2 will have a high degree of co-citedness with the citing article.
- Move three will only contain few or none citations, but documents that are cited in this move, will have the highest number of co-citations with the citing article.

2 Method

2.1 Source Articles

The empirical basis was obtained by analyzing a number of research articles, noting down the location of the references, and counting the number of co-citations to the source article and the cited document.

In order to get as clear a picture as possible, it has been a priority to choose the source articles in such a way as to minimize the amount of uncontrollable variables. At the same time I set up certain requirements that severely limited the amount of suitable source articles.

- The source articles should be available in full text, so that I can examine the move structure, and pinpoint the location of references.
- The articles should be published in journals that are indexed in a major citation index (SCI, SSCI, A&HCI or Scopus; Google Scholar does not allow for co-citation studies). Though it would be possible to use one of these indexes to perform co-citation analysis on journals that are not included in the index, it would be more tedious and the anticipated large part of the co-citations that comes from later articles in the same journal would not be included in the study.
- The references should be numbered, as it would be too time-consuming to sort them out myself.
- The articles must have enough citations to make the number of co-citations meaningful.
- The articles should be contemporary yet old enough to allow for a sufficient amount of citations and co-citations to accumulate.
The articles should be structured along the same lines, so that they are comparable.

It was found that *Journal of the American Chemical Society* is an ideal periodical for this purpose. It is available in full text from the Royal Library; it uses numerical references ordered after first appearance in the text (American Chemical Society 2011); it is indexed in SCI; it is American, neutralizing possible bias in SCI due to nationality and language (Seglen 1997); and it is highly cited, receiving 351813 citations in 2009 (Thomson Reuters 2011).

The 30 highest cited articles from 2000 were chosen as source articles, though shorter pieces defined as articles with only one or two pages were discarded. Year 2000 seems a reasonable choice since *Journal of the American Chemical Society* has a Cited Half-life of 7.3 (Thomson Reuters 2011). The highest cited articles were chosen in order to achieve as reliable a data sample as possible. I can see no reason to suspect that the highest cited articles should be differently organized than the typical paper.

The source articles and citations to both the articles and the cited documents were retrieved in SCI via Dialog, using the following method:

```
S1 56041 JN='JOURNAL OF THE AMERICAN CHEMICAL SOCIETY'
? e cr=gronbeck h
Ref Items Index-term
[...] E34 1 CR=GRONBECK H, 1999, V122, P3839, J AM CHEM SOC
E35 1 CR=GRONBECK H, 2000, P3839, J AM CHEM SOC
E36 1 CR=GRONBECK H, 2000, V112, P15940, J PHYS CHEM C
E37 4 CR=GRONBECK H, 2000, V122, P2839, J AM CHEM SOC
E38 273 CR=GRONBECK H, 2000, V122, P3839, J AM CHEM SOC
E39 3 CR=GRONBECK H, 2000, V122, P3843, J AM CHEM SOC
E40 1 CR=GRONBECK H, 2000, V122, P3939, J AM CHEM SOC
E41 93 CR=GRONBECK H, 2000, V262, P1, CHEM PHYS
[...]?
```

```
s e35,e37:e40

1 CR=GRONBECK H, 2000, P3839, J AM CHEM SOC
281 CR=GRONBECK H, 2000, V122, P2839, J AM CHEM SOC
S2 282 E35,E37:E40
```

```
s s1 and au=gronbeck? and py=2000

56041 S1
989467 PY=2000
S3 1 S1 AND AU=GRONBECK? AND PY=2000
```

? . rank cr
The results were then copy-pasted into excel, and the references numbered according to the sequence in which they were cited. Unfortunately SCI has chosen to alphabetize the references even though they are numbered in the original article, which is somewhat puzzling considering the preference the founder of SCI ostensibly has for numbered references (Garfield & Small 1997).

Only the first occurrence of each citation was recorded, which is unfortunate since it would no doubt change the picture if all occurrences of the cited documents were included. However, repetitions of citations were excluded from the study due to the excessively laborious and time-consuming work it would have involved.
2.2 Determining the Rhetorical Moves

Determining the moves in the introduction of an article is a subjective process, though not entirely haphazard as can be seen by the reasonably high inter-rater stability achieved by Crookes (1986).

Often, the moves can be identified by certain signal-words or phrases, such as “In this study...” indicating move 3, or “however”, which is a fairly typical way to start move 2. In other instances only an expert of the specialty can discern the shift of move with certainty. Crookes (1986) gives an example of this:

... These methods are useful for comparing intact filaments in different tissues. They do not indicate whether individual filament forming polypeptides (FFPs) have significantly large regions of related aminoacid sequences.

In this case, even a close reading by a non-specialist may miss the fact that the second sentence is raising a gap. It is necessary to have an understanding of the relationship between aminoacid sequences and FFPs (Crookes 1986, p. 66).

Sometimes the signal-words are deceptive, for example the phrase ‘here we ...’ is recurrent in move 3, but the sentence “but here we will describe the characterization of a molecule...” (From article #25, see appendix B) actually occurs long before move 3 starts. I take this to be read as ‘in this connection the focus will be on ...’ as the statement is followed by background information and not information on the current study. Thus I coded this sentence as a move 1.

Another case is a phrase like “it is only recently that there have been reports ...” (From article #30), adversative words are commonly used in the opening of move 2 (Swales 1990, p. 154), but here it is used to emphasize currency, and thus belongs in move 1.

As an aside it can be noted that these examples show that it would be tricky to create a term-based algorithm that could recognize the move-structure of an article.

Initially the move-structure of the introductions where determined at sentence-level, which is a practice that has been used in earlier studies (e.g. Crookes 1986). The actual assignments of moves to sentences are presented in appendix B (p. 56). As it turned out, this method was not ideal for my purpose. Sometimes there are several cycles between the moves, so that the structure is for example 1-2-1-2-1-2-3 instead of simply 1-2-3. In fact most of the source articles cycle several times through move 1 and move 2, the pattern being that a move 1 sentence with a statement about current knowledge quoting a reference, is followed by a move 2 sentence qualifying this statement without citing a reference, such as this example from article #19:

The polymeric sp carbon allotrope is often referred to as “carbyne”, and numerous syntheses have been claimed.\(^5\)\(^6\) [Move 1]

However, most samples are intractable or difficult to characterize. [Move 2]

This pattern is repeated several times, and references are predominantly placed in move 1 sentences. Since references are rare in move 3, the result is that the vast majority of references are located in move 1 sentences, making it hard to make
definite conclusions on the co-citedness of the few citations in move 2 and move 3. So I chose to supplement the sentence-level coding with paragraph-level coding and algorithmic coding.

The paragraph-level coding is done by determining the main rhetorical move for each paragraph in the introduction. Paragraphs that mainly describe current knowledge are coded as move 1, paragraphs with gap indications are coded as move 2, and paragraphs introducing the research of the current article are coded as move 3.

In practice this was done ‘backwards’ by coding paragraphs that mention the research in question as move 3, other paragraphs that in my opinion criticize previous research or in other ways point toward a potential research area are coded as move 2, and the remaining paragraphs are coded as move 1. The assignment of moves on the level of paragraphs can be seen in Table 2 (p. 42).

The algorithmic coding is simply the relative position of the citation in the introduction, calculated by paragraph. The introductions in the study contain from 1 to 10 paragraphs, and the tripartitions were assigned according to Table 1.

Table 1. Tripartitions for the algorithmic assignment of move.

<table>
<thead>
<tr>
<th>No. of paragraphs in introduction</th>
<th>Move 1</th>
<th>Move 2</th>
<th>Move 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Void</td>
<td>Void</td>
<td>Void</td>
</tr>
<tr>
<td>2</td>
<td>§ 1</td>
<td>-</td>
<td>§ 2</td>
</tr>
<tr>
<td>3</td>
<td>§ 1</td>
<td>§ 2</td>
<td>§ 3</td>
</tr>
<tr>
<td>4</td>
<td>§ 1</td>
<td>§ 2-3</td>
<td>§ 4</td>
</tr>
<tr>
<td>5</td>
<td>§ 1-2</td>
<td>§ 3</td>
<td>§ 4-5</td>
</tr>
<tr>
<td>6</td>
<td>§ 1-2</td>
<td>§ 3-4</td>
<td>§ 5-6</td>
</tr>
<tr>
<td>7</td>
<td>§ 1-2</td>
<td>§ 3-5</td>
<td>§ 6-7</td>
</tr>
<tr>
<td>8</td>
<td>§ 1-3</td>
<td>§ 4-5</td>
<td>§ 6-8</td>
</tr>
<tr>
<td>9</td>
<td>§ 1-3</td>
<td>§ 4-6</td>
<td>§ 7-9</td>
</tr>
<tr>
<td>10</td>
<td>§ 1-3</td>
<td>§ 4-7</td>
<td>§ 8-10</td>
</tr>
</tbody>
</table>

2.3 Comparing Documents

To truly establish the semantic relationship between two documents, one should have domain experts read and compare the two texts in question, but I had to limit myself to algorithmic methods, which can at best estimate the probability that two documents are semantically related. The available methods include: comparing the metadata of the documents, comparing the text of the documents and using citation analysis.

Metadata. The source articles are provided with subject categories, which are assigned by indexers to each journal in SCI. All articles from a specific journal will thus have the same subject categories. This is clearly inadequate for my purpose.

The only alternative semantic metadata available for the source articles is KeyWord Plus, which are algorithmically drawn from the titles in the reference list of the article. Though KeyWord Plus might enhance information retrieval, it has not been used much for similarity studies. Furthermore there are only 2 to 10 KeyWord
Plus for each article, whereas the average number of references is over 50, giving a more solid base for the calculations.

**Full Text.** Comparing the entire text of two articles using for example the vector space model is a well-tried and efficient way of estimating similarity. It would also allow me to test documents that are less-cited. Unfortunately, this method would require me to track down full-text versions of the 1700 cited documents in the sample, which would have been an insurmountable task.

**Citation Analysis.** Evaluating the likeness of documents by citation analysis is usually done by counting either bibliographic couplings or co-citations, of the two methods co-citations were deemed to be the better choice, as it is recommended by several citation analysts (e.g. Small & Griffith 1974; Egghe & Rosseau 1990; White 2004).

### 2.4 Co-citations

It appears [sic] that an interpretation of the significance of strong co-citation links must rely both on the notion of subject similarity and on the association or co-occurrence of ideas.” (Small 1973).

The purport of citations is not as straightforward as one might think. In the theory-section later I will present views on the different meanings assigned to citations, and by implication to co-citations. But here I will give a short practical reasoning on why co-citations are expected to signify semantic relatedness.

A citation is the author’s declaration that the cited document is relevant in some way to the text area (word, phrase or claim) where the reference is located. On the assumptions that the author is a specialist in the subject on which he publishes a paper, and furthermore that the author is the most salient expert of his own text, the conclusion is that something in the cited document has some kind of relevance to the part of the citing document that contains the reference, unless we ascribe sinister and hidden motivations to the author.

Co-citedness means that two documents are both cited by a third, later document, since both of the cited documents have a relevance-link to the citing document; it follows that they might also be related to each other. We cannot be sure about this, since they might be cited from two different unrelated parts of the citing document, yet it is reasonable to assume that the probability that two documents are semantically related is higher the more co-citations they have.

The weight of a co-citation link between two documents depends both on the number of co-citations and the number of citations each of the two documents have received. In the example shown in Fig. 2 the co-citation link between documents No’s 2 and 4 is very strong because they have three co-citations (by No’s 6, 7 and 11). The link between documents No’s 6 and 7 is weaker since they only have one co-citation, but their bond it is nonetheless stronger than the co-citedness of documents No’s 4 and 5, for even though they also have one co-citation, they each have three citations, making the proportion of co-citations to citations less.
Fig. 2. Illustration of co-citations and bibliographic couplings. Numbered circles represent documents, arrows points from a citing to a cited document (Cawkell 1974).

Vector Space Model. When co-citations were introduced independently by Small (1973) and Marshakova-Shaivekovich (1973), they both noted that the degree of co-citedness of two documents is not simply the amount of co-citations, but that this number ought to be set in relation to the amount of citations the documents have received, though a precise algorithm was not presented. I have decided to use the vector space model (Salton et al. 1975) for this purpose. This model sees documents as vectors in a multidimensional space. The number of dimensions corresponds to the number of different terms in the base, and the direction and length of each vector is defined by the totality of terms in the document. Similarity of two documents is then defined as the cosine to the angle between their vectors.

In this study terms are substituted with citations. Each reference is treated as a document, and each citation to the reference is a term in this document, the term frequency is thus either 1 or 0 and the document length is equal to the total number of citations given to the reference. The 30 source documents are treated as queries, and the resulting algorithm used to calculate the similarity between a source document $i$ and a cited document $j$ is:

$$sim_{ij} = \frac{cocit_{ij}}{\sqrt{c_i} \cdot \sqrt{c_j}}$$

Where $cocit_{ij}$ is the number of co-citations between the two documents, and $c_i$ and $c_j$ are the number of citations to the cited document and the source document.
respectively. This formula cannot handle documents without citations, but that will never be the case anyway in this study.

To continue the examples from Fig. 2 above, the similarity score for documents No’s 2 and 4 would be: ————, for No’s 6 and 7 the score is ————, and the similarity between No’s 4 and 5 only score ————. Similarity scores calculated this way will range from 0 (no co-citations) to 1 (the two documents are always cited together).

The vector based similarity is sensitive when the citation rates are low; to have stable scores it requires that the two documents have received an ample amount of citations. The source documents in this study have been chosen because they are highly cited and have at least 224 citations, but some of the cited documents have only received the single citation from the citing source article. It follows that the two articles cannot be co-cited in these cases, but it would be rash to conclude that they are not similar based on that.

Therefore I decided to discount the lowest-cited documents from the analysis; I set the cut-off at the median, which is 81 citations, meaning that only the 849 cited documents that have at least 81 citations are included in the analysis.

2.5 Potential Methodical Problems

**Age of References.** If there is chronological difference on two documents, co-citations can only occur after the publication of the newest document. Since I modify the raw co-citation counts with the number of citations received, this means that older references are punished, the citations they gain before the publication of the source document in 2000 weigh them down even though they cannot gain any co-citations in this period. This problem has an impact on the cited documents in the method section, which generally are older and more cited than the rest of the references.

The solution to this problem would be to limit the Dialog searches to 2000-2011. This has unfortunately not been practically possible in this study, but the problem is mitigated because searches were limited to Dialog base 34, which only goes back to 1990.

**Co-citation Analysis on Cited Documents.** This study uses co-citation analysis on a given document and the document it cites. Documents with a direct reference/citation link generally have more co-citations than non-linked documents (Small 1973). But since all the co-citation counts in this study are between documents sharing such a link, they are all biased in the same way. The co-citation counts will therefore be expected to be higher than usual, but that should not distort the picture of the relative relatedness of the documents. On the contrary, the higher numbers might conceivably by sheer numbers giving a more solid basis for conclusions.

**Self Citations.** Self citations are a general problem in citation analyses, some researchers claim that they are inherently biased, produced by self-promotion and egotism (e.g. Hyland 2000, p. 23), others see them as natural and helpful given that the author knows his own works thoroughly, and often the most pertinent earlier research will be the author’s own work (White 2001; Swales 2004, p. 84). However,
whether one or the other side is correct, it would be an overwhelming task to sort out all the self-citations since most of the articles in Chemistry are written by multiple authors. For example, the 30 source articles in the sample of this study had from 2 to 13 authors averaging 5 (Table 2, p. 42). Since SCI only list the first author of the references, it would be necessary to look up all the 1700 cited references to check all the authors, not all of these cited documents will be included in SCI, so this process would entail a long and arduous detective work.

It is also conceivable that there might be bias in the form of references to the same institution or journal, and that these biased references might generally be placed in the same place of the source articles, but this possible source of bias were not examined.

Science Citation Index (SCI). This study is limited to SCI’s database. Citing documents that are not included in SCI will therefore not contribute to the citation and co-citation counts. This can be a problem as SCI does not include books, which might introduce a distortion in the study, since books have a different citation practice from articles (Smith 1981, p. 92).

On an equal note, some citing articles will be lost due to errors in the database. However, errors due to faulty posts will probably be randomly spread across the articles.

3 Theory

This study explores citations in the context of genre; therefore the following section will explain some of the most important issues in genre analysis and citation theory. First there will be some details of the concept of genre, with a short description of different ways to see this concept. This is followed by an account of the scientific community in the light of discourse communities as the primary structuring units of science. An analysis of the research article as a genre with specific focus on the introduction is then followed by a detailed historical and theoretical description of citations ending with a proposal for a genre-based theory of the functions and implications of citations in research articles.

3.1 Genre

On the face of it, it seems a simple task to ascribe genre to a document. However, not only is it sometimes difficult to pinpoint which genre a text belong to, but the division of genres is not given a priori. Genres can be divided in larger or smaller groups (Adams Smith 1987), and over time genres evolve, and new genres are created. Even so, the concept of genre can be an efficient tool to understand the rhetorical workings of a document, if one remembers that the classification of genre always will be subjective.

Swales (1990, p. 58) had a clear definition of genre: “A genre comprises a class of communicative events, the members of which share some set of communicative purposes”, but he has since re-evaluated his position (Swales 2004, p. 62) and realized that a rigid definition can be counter-productive for the understanding of
genre, and that it can result in blind angles instead of new knowledge. This is why,
instead of defining genre, he illustrates the concept in a very flexible way with a
number of metaphors (Swales 2004, p. 61-68):

As a frame to be filled as needed, a genre is the environment where communication
takes place. As such it helps shape the thoughts of authors and readers, but the genre
itself is a frame, and though genre knowledge is necessary, it is not sufficient for
meaningful communication.

As a standard to be followed or diverted from, a genre both constrains and offers
choice. Much like language, a genre has no absolute rule, but a total disregard of the
standard would incapacitate communication and be considered inappropriate both
rhetorically, but in some occasions also socially.

As a species that evolves and multiplies, genres cannot be definitely defined, it is
always possible to divide a genre into new sub-genres. New genres appear when a
need for them develops. Genres evolve never being absolutely stable, and they will
acquire new functions as needed.

As a family, a genre comprises instances that have mutual resemblances but no
common characteristic. The genre will consist both of typical members and outliers,
and some of the outliers will have nothing in common. Our notion of which
documents belong to a genre is not guided by a set of identifiers, but by a sense of
resemblance with prototypical members of the genre.

As an institution, a genre is more than the documents it comprises, and more than
its users. A genre has a life of its own, and it exists in a network with other genres and
institutions. The participants of a genre, both authors and readers, take on roles within
the genre.

Genre Analysis. Genre can be defined from the text in itself (style, lexis, topics),
from the context (the communicative process), or using the generic labeling of the
user group. Whatever method is employed, it is advisable to repurpose the genre after
the initial classification (Swales 2004, p. 73). As we shall see later, Swales’ analysis
of the research article genre is based on the third of these methods, using the generic
labeling of academia as a starting point, and redefining the classification to reflect
common practice.

Genre analysis can be performed on different levels; if we want to use it to predict
the importance of references, we must look beyond superficial characteristics as
lexical density and complexity (e.g. Biber & Finegan 1986), or linguistic elements
surrounding the reference (e.g. Swales 1981). Instead, we should look at the rhetorical
organization of the text (Widdowson 1979, p. 60; Adams Smith 1987), in order to do
this we must look at the cognitive, rhetorical and linguistic aspects (Swales 1981, p.
10). Of course, the superficialities are elements that can help us to discern the shift of
rhetorical moves.

It is not the purpose of this study to perform a genre analysis on the research
article; instead, the results of the genre analyses of research articles performed by
John Swales (1981; 1990; 2004) has helped me to form hypotheses on the import of
the location of references, hypotheses that are the basis of this study.
3.2 Discourse Communities

A genre can only be fully understood if one also studies the users of the genre, i.e. the discourse community that employs the genre. A discourse community is a group of people primarily bound together through communicative bonds. The community has a distinct language, members show their affiliation with the group through communicative acts, and new members are initiated by learning the particulars of the discourse (Swales 1990).

In this context we are interested in discourse communities in science, meaning groups of scientists who form discourse communities. That such communities exist can for example be discerned by looking at patterns of citations (e.g. Nicolaisen 2004).

In science communication is a defining characteristic of communities. The community is only a community because of the discourse, and the discourse can only exist within a community (Hyland 2000, p. 10).

The communication of a discourse community takes the form of one or more genres and the choice of genre depends on the communicative purpose.

Genres that are common to different discourse communities, such as the research article, can be expected to differ in various ways across the communities.

To establish the background information needed to explain the workings of research articles, the next two sections will describe the scientific community and the division of science into discourse communities.

3.3 The Scientific Community

The totality of active scientists can be referred to as the Scientific Community, and can be seen as a large and loosely-tied discourse community in itself. The scientific community has a common goal: to increase scientific knowledge, and holds to some general ideals such as objectivity and rationality. Sociologists of science have given diverse explanations of the nature of Science.

CUDOS. The normative theory of science revolves around the idea that the actions of scientists are regulated by tacit norms. Norms are not descriptions of actual behavior, nor are they official rules, rather norms are values shared by a majority of scientists, who will encourage behavior that is in accordance with these norms. The ethos of science was formulated by Merton (1942) as four norms: Communism, Universalism, Disinterestedness and Organized Skepticism (CUDOS).

The gist of CUDOS is that scientists should not have personal interests vested in research, scientific claims should be judged by their merits alone, and not influenced by the gender, faith, nationality etc. of the author, but all scientific claims should be rigorously tested before acceptance and possible incorporation into codified knowledge. Science belongs to all of humanity, results should be published for the benefit of all, and the discoverer of new knowledge does not have any rights over said knowledge, except for the right to be credited when other scientists use his discovery.

In the normative view, science is implicitly seen as one large discourse community sharing common ideals and standards. This view has since been challenged.
Paradigms. Kuhn (1970) sees science as dependent on the overall world-view. The current paradigm is decided by the concurrence of scientists. It is impossible to prove that one paradigm is superior to the other, since all argumentation is based on the rules within a paradigm itself. This does not mean that the paradigm is solely decided by social intercourse separated from reality, but rather that reality allows for multiple possible paradigms.

Paradigms are ‘the models of science’, a pattern that demonstrates how to perform science. Kuhn also uses the term paradigm for accepted theories that decide which questions can be asked and what kind of answers will be acceptable. By extension ‘paradigm’ is used for the group of adherents of such a world-view.

Paradigms are not defined as a set of dictums but by exemplars. This means that the concept of paradigms is closely connected to the concept of genres, which are also defined by exemplars rather than by set rules. The acquisition by a student of a scientific paradigm is simultaneous and cooperative with the acquisition of the necessary genres of the discourse community.

The influence goes both ways: paradigms are decided by the consensus of scientists, but on the other hand, scientists cannot help but think according to the paradigm. In fact few scientists are capable of changing paradigms, and when paradigms change in periods of revolutionary science, it mostly happens slowly by young scientists adopting the new paradigm, and old scientists dying out (Kuhn 1970).

The smaller units of science, the discourse communities of specialties and research traditions are also defined by the paradigm they follow. To study a specific specialty in science is to belong to a scientific discourse community. It does not simply entail to find the answer to a certain problem, but is a way of life (Geertz 1983, p. 155). A paradigm influences the entire world-view, and is thus not easily changed for another. Indeed, still according to Geertz, scientists of different discourse communities live in different worlds that cannot be integrated, and the best one can hope for is communication across the borders. Kuhn (1970, p. 112) also notes that misunderstandings are likely between different paradigms.

The subjectivism of science that the concept of paradigms suggests does not entail that science is just a matter of opinion. Observations and empirical evidence are still at the basis of scientific research, but the observations allow several plausible explanations (Kuhn 1970, p. 4; Hyland 2000, p. 20).

Social Constructivism. The social constructivist point of view is that knowledge is a construction that is the result of social processes. Research papers are more than reports they are actively used to shape academic knowledge through discourse (Hyland 2000).

Separating science entirely from the supposed ‘reality’, Bruno LaTour (1987) postulates that there is no qualitatively difference between fact and fiction (p. 42), that it all boils down to who wins the debate. Science is a battle of ideas, and research articles are attacks that employ references, stratification and graphs to persuade opponents of the validity of the author’s claim. Fact and fiction in a specialty are decided by the specialists, who stay specialists by constant training in the disputes with each other, therefore there is no isolated specialists.

Paradoxically, such claims as “Rational argument is a social matter, governed by disciplinary norms and oriented to achieving an intersubjective consensus through
persuasive means” (Hyland 2000, p. 8) are defended using the very rational arguments that are undermined by the claim. It might be wise to remember that the sociology of science in studying science also studies the sociology of science itself (Merton 1976, p. 56).

Despite the obvious disagreements there is some likeness between social constructivism and the two abovementioned theories: Research articles are vehicles of the claims of the author, the purpose of the article is ultimately to add to the scientific knowledge, and therefore the article should convince readers of its validity. Specialties exist through the discourse between specialists, and truth is not absolute. The major divide between social constructivism and the other approaches seems to be of a moral kind: Are scientists pursuing truth or influence?

### 3.4 Research Traditions and Specialties

Grand rubrics like "Natural Science," "Biological Science," "Social Science," and "The Humanities" have their uses ... the sorts of work conducted under any one of them do show some general resemblances to one another and some genuine differences from the sorts that are conducted under the others. ... But when these rubrics are taken to be a borders-and-territories map of modern intellectual life, or, worse, a Linnaean catalogue into which to classify scholarly species, they merely block from view what is really going on out there where men and women are thinking about things and writing down what it is they think (Geertz 1983, p. 7).

Traditionally Science is split in a tripartition between Natural Science, Social Science and Humanities. That this is an arbitrary division and other meaningful divisions are possible can be seen by such alternative dichotomous splits as ‘hard vs. soft’, ‘dry vs. wet’ (Schneider 2011) and ‘urban vs. rural’ science (Swales 2004, p. 15).

In spite of the Geertz quotation above, it can be enlightening to look at the primary units of science. According to Nicolaisen (2004) the primary units of Science are both specialties and research traditions. These primary units are small enough that everyone knows everyone, and they form closed groups where most communication is internal.

Indeed, when we get down to the substance of things ... most effective academic communities are not that much larger than most peasant villages and just about as ingrown. Even some entire disciplines fit this pattern: it is still true, apparently, that just about every creative mathematician ... knows about every other one, and the interaction ... among them would make a Zulu proud. To some extent the same thing seems to be true of plasma physicists, psycholinguists, Renaissance scholars, and a number of other of what have come to be called, adapting Boyle's older phrase, “invisible colleges.” From such units, intellectual villages if you will, convergent data can be gathered, for the relations among the inhabitants are typically not merely intellectual, but political, moral, and broadly personal (these days, increasingly marital) as well (Geertz 1983, s. 157).

At a more formal level, these differences between discourse communities can be seen in the variation of vocabulary, journals and conferences that scientists
communicate through, and which in some instances can result in different groups researching in the same area but oblivious to the existence of each other (White 2004, p. 91).

The reason to operate with both specialties and research traditions as primary units is that they intersect, so that a research tradition can have adherents in several specialties, and a specialty can be studied by several research traditions. Both research traditions and specialties have distinctive practices when it comes to the overall structure of research articles (Hjørland 2002), and the number of references and how they are incorporated in the text (Hyland 2000).

The primary units made by the intersecting research traditions and specialties could be called discourse communities of science. Nicolaisen (2004) concludes that communication is markedly stronger within such primary units than across the borders, which corresponds with the tendency to misunderstandings between paradigms noted by Kuhn.

3.5 The Research Article

The central genre of science is the research article. Though it is not the primary document type in all areas of academia, it is an important genre everywhere, and the primary genre for a majority of research areas (Swales 2004, p. 13). Its importance lies both as a formal communication channel, but also as a major element in the codification of scientific knowledge.

Other formal genres in and around academia include popularizations (e.g. articles in New Scientist or Scientific American), books, theses, conference papers, and book reviews. These genres vary in overall organization, textual style, number and kinds of references etc. The following paragraphs concentrate on journal articles, and should not be taken as generalizable to other academic genres.

**Purpose of the Research Article.** The grand ideal purpose of the research article is arguably to contribute to the academic discourse, and to hopefully add a tiny morsel to the collective human knowledge, which Popper (1978) called world 3. That this is not always the case can be seen by the fact that the majority of published papers never receive any citations (Cole 2000; Swales 2004, p. 246).

Apart from this ideal purpose, research articles have a more practical purpose for scientists as a form of symbolic capital. Published articles in notable journals give recognition among peers and may result in better career options (Swales 2004, p. 246), uncited articles might not result in positive recognition, but not publishing will in most places have negative implications for one’s scientific career. Furthermore, in some countries (e.g. Denmark and Norway) publications are translated into points that can affect the economy of individual institutes and thereby the research opportunities of scientists.

A third purpose for publication of research articles is along the lines of ‘homo ludens’: that the scientist enjoys the challenge of getting an article published in an exclusive journal (Swales 2004, p. 246). This corresponds with puzzle solving as standard behavior during periods of normal science according to Kuhn (1970), the
puzzle is only really solved when the solution is recognized by one’s peers (i.e. is published).

**Genre or Macrogenre?** Widdowson (1983) describes the research article as a macrogenre. Dudley-Evans (1987) calls it a genre in itself but splits it into subgenres, so that the effect is somewhat the same:

A genre may have a number of sub-genres. The research article genre is very broad, and can be broken down into a number of sub-genres such as the survey paper, the conference paper, research notes (a shorter form of the article reporting important results but with little comment) and the letter. Genre analysis may be used as a classificatory system, revealing the essential differences both between the genre studied and other genres and also between the various sub-genres.

Swales (1990, p. 175) questions that the research article is a macrogenre, but elsewhere he also subdivides it into several genres in a similar way as Dudley-Evans. Swales’ division is in *theory pieces, review articles, data-based research articles and shorter communications* (Swales 2004, 207).

**Subdivisions of the Research Article According to Swales.** The data-based research article is the classical research article. SCI does not discern between this and theory pieces, which are common in Astrophysics, Theoretical Physics, Mathematics, Economics and Linguistics. Theory pieces usually have an overall structure going from general to specific, thus not complying to the hour-glass model (Swales 2004, p. 207), and they can have different citation-practices than experimental articles (Chubin & Moitra 1975).

Review articles, which give a survey of the research in a specific field, are a well-established category that several journals and SCI use. Swales describes it as having more references, and often having only one author. He hints that the move/step structure is atypical for the review article (Swales 2004, p. 208-211).

Letters, which also are an independent document type in SCI, are regarded as analogous in structure to the full-length articles by Swales (2004, p. 234), though they possibly have a different citation practice (Chubin & Moitra 1975). Swales’ shorter communications are even shorter than letters: conference preprints, squibs, short notes etc., and seems to be a conglomerate grouping of leftovers. Apart from these four subdivisions, Swales mentions rebuttals (Swales 2004, p. 65) and meta-analyses (p. 273), the latter being a special genre within medicine merging together the results of multiple research articles. The IMRAD structure is common in data-based research articles, but is not used in theory pieces or review articles. It is not clear whether Swales believes that the CARS model is applicable to these two subdivisions or only the data-based research articles, but he does state that the structure is different though comparable in shorter communications.

**Dimensions Important for the Research Article.** I will in this section focus on the data-based research article, as it is the subject of my study. The current study tries to empirically confirm or disprove a hypothesis that the location of a reference in a research article can be used to predict the semantic relation of the citing and the cited document. However, whether the hypothesis holds or not for my sample, the results are only valid for the purposely very limited area that this study deals with. The study is limited in order to diminish the amount of uncontrollable variables as much as
possible. Here I wish to explore the kinds of dimensions that might change the picture.

That there is a difference in reference practice between hard and soft science was shown already 40 years ago (Price 1970), also corroborated by Small (1982). Chubin & Moitra (1975) have argued that reference practice varies over specialties, and Jacoby (1987) exemplifies this with the discipline of literary research.

Swales (2004, p. 244-245) showed that there are differences in the structure of research article introductions in different languages, specifically that some non-English languages use more relaxed rhetoric in the introduction, which might be due to a less competitive research space. Differences in citation practice between languages are affirmed by Chubin & Moitra (1975) and MacRoberts & MacRoberts (1989), who also asserted differences in citation practice between national cultures.

Everything changes over time including research articles, from the modest beginnings in the 17th Century to today, research articles have changed in length, in the amount and type of references, style of language, type of illustrations and subdivisions of the text (Swales 1990, p. 110-116).

Lastly one should remember that the possible rules for citation-practice that can be established will always be general rules and not absolute (Hyland 2000). As Swales (1990, p. 145) writes: “for, even more than syntax, discourse is a phenomenon of propensities. Discourse generalizations are permeable to exceptions”, though he also states that these idiosyncrasies are less marked in humanities than in natural science, and that research articles are more stable than most genres, since they are peer-reviewed (Swales 2004, p. 18).

To recapitulate, citation practice is variable across time periods, subjects, language and publication types (MacRoberts & MacRoberts 1989). The dimensions where it is expectable that there are differences in rhetorical organization and citational practice are the following:

- Specialty.
- Language.
- National culture
- Period.
- Idiosyncrasies.

and possible also:
- Research tradition.
- Subgenres, such as case-reports and the clinical trials in Medicine.

The point is not that there are no common rules for all research articles, rather, the point is that we cannot assume that principles proven for say, American chemistry articles published in 2000, can automatically be expected to apply for all other research articles. There are many resemblances in research articles across all dimensions (Swales 2004, p. 216), however, reasonable diverse samples must be studied across all these variables, before general appliance can be assumed.

Authorship. Research articles often have multiple authors, especially in the natural sciences it is not uncommon to see ten or more co-authors, and this phenomenon makes it likely that the research article is a more stable and less idiosyncratic genre than single-authored genres (e.g. the letter or the book review).

The long process of writing a research article enhances this tendency, bringing in co-creators beyond the official author(s). Most authors of scientific papers confer with
colleagues and presents future papers at conferences, using the feedback to improve their articles (Hyland 2000, p. 12).

Furthermore, research articles are published in peer-reviewed journals; as such there are added restrictions and normalizations of the paper. Firstly, all journals have standards that petitioners must follow: reference-system, font-size, structure etc. Secondly, the author(s) of a research article must consider the gatekeepers (Hyland 2000, p. 18) - the editors and reviewers of the journal, who decides which articles get published. Thirdly, in most cases the reviewers require revisions of the text, so that the final published version of the article is jointly produced by the author(s) and the reviewers.

The author(s) also has to keep the potential readers in mind. They are a small group of specialists in the field, who know the background literature and the schema, and who the author(s) must try to persuade that this current paper belongs among the accepted scientific knowledge.

In the rest of this paper, when I write about the authorship of generic articles I refer to the author(s) in the singular masculine for ease of reading.

History of the Research Article. The Research article is an academic genre, and as such it is constructed along the lines of the general conventions for scientific communication. Scientific communication is on the face of it almost as the absence of genre, only showing ‘reality as it is’ using neutral terms, impersonal style and logical reasoning (Hyland 2000). But as Swales (1990, p. 110-176) demonstrates, research articles are constructed through a set of norms that guide organization and language. This set of norms seems natural to us, because we have learned to see science through these norms, but, as all other norms, they have evolved over time.

Already in the 17th Century norms of scientific honesty, language and argumentations are expressed by people like Francis Bacon and Robert Boyle. These norms were developed in order to secure replicability and rational reasoning, but also to make the report of the experiment more convincing (Swales 1990, p. 111-112). The norms from that time are still recognizable today, and an obvious conclusion would be that they have been standards for the last 3 or 4 centuries. That is not quite how it is, scientific norms and practice change like everything else.

The research article had predecessors in the form of monographs, anthologies and scholarly correspondence, genres that were merged into the genre of journal articles when the first scientific journals were published in 1665. In time the research article gained its own form, and around the year 1800 the research article had acquired many of its modern characteristics. That is not to say that the process has ended and that research articles do not develop anymore, indeed they still to change and will do so continually.

Description of the Research Article. Research articles are written by specialists for specialists, and the specialties are often very specific. Much information is implicit, and the background knowledge that is expected of the reader can be so comprehensive that it can be extremely difficult or outright impossible for an outsider, including scientists of other specialties, to decipher (Kuhn 1970, p. 55). Another characteristic is that a research article is often a condensed report of previous research (Swales 2004, p. 20), perhaps concentrating months of work into a few pages, making the resulting paper very information-rich.
The language of research articles is characterized by the author’s wish to persuade the reader of the value of his claims, i.e. that the claims are original, important and true. In order to persuade the reader, the rhetoric will be in accordance with the assumptions, theories, methods, and knowledge of the paradigm, and subscribe to epistemological and social conventions of the discourse community (Hyland 2000, p. 13-14).

**Ambiguities.** One might imagine that the desire to persuade the reader would lead the author to use pompous or conceited language, when this seldom is the case; it is because the author wishes to limit the possibility of future criticisms (Merton 1976, p. 62). This is why a normal structure of research articles is the combination of argumentation for the value of a claim and hedges to limit the validity of the claim (Hyland 2000, p. 13). Also the references to earlier research is of an ambiguous nature, on one hand the author must place his study in the acknowledged research tradition, on the other hand it is necessary to find some sort of gap of knowledge in order to justify the current research, an ambiguity that was already pointed out by Merton (1957). These ambiguities can be an obstacle if one tries to classify references by the surrounding text, as an apparent accept or critique in the text possibly has rhetorical reasons rather than being meant as an opinion of the cited document.

All the above observations regarding the language of research articles have implications for references as well because references are an integrated part of the rhetoric of a text (Gilbert & Mulkay 1984; Jingfu 1987). A reference is a link to another text, whether it is on the level of document, paragraph, sentence, phrase or word, and this connection is arbitrarily made by the citing author (Leydesdorff 1998), and is thus a part of the deliberate communication between writer and reader. To understand the purpose of references, one must understand the entire communicative process in the scientific text. References cannot be understood separately from the accompanying text.

### 3.6 The Organization of the Research Article

The following description is based on the IMRAD structure (i.e. Introduction, Methods, Results And Discussion) which is fairly common throughout research articles in the natural sciences. But many of the following descriptions also apply on a general level to other research articles.

The overall organization of the research article goes from general to specific in the beginning, and from specific to general in the end, as illustrated in the hour-glass model (Fig. 1, p. 8). The introduction starts on a broad level, gradually zooming in, and ends up focusing on the research gap, followed by an introduction of the current research. The method and results sections keep their focus on the current study, but in the discussion the perspective becomes more general again. The difference between the introduction and the discussion apart from the direction of the change of scope, is that in the introduction the existing research is in focus, explaining why the current study is justified. Whereas in the discussion the results of the current study are in focus, and if references are included, it is to add perspective (Swales 2004, p. 234-235). This is why the introduction holds more references than the discussion.
The hour-glass model is widely accepted, though corrections have been proposed. For example Jingfu (1987, p. 103) suggests that the discussion section can be repetitive and go from specific to general multiple times; once for each research question.

**The Introduction.** Most references are located in the introduction (e.g. Voos & Dagaev 1976; Jacoby 1987). This complies with the twofold demand to place one’s research in the proper research tradition and to argue that one’s research is original (Hyland 2000, p. 21). What better way to do this than to review earlier studies and describe what they have done and what they have not done, then compare with one’s own research and how this new study is doing what the earlier studies failed to do.

Cano (1989) agrees that most citations are given in the introductions, but believes that the majority of these citations are perfunctory. Opposing this, Paul (2000) states that the introductions are important parts of research articles, and that references located here will be central to the citing article, especially if it is referred to multiple times in the introduction, and she ascribes this to the fact that introductions are one of the most read sections of research articles, third only to abstracts and conclusions. Contrary to this view Jingfu (1987, p. 107) writes that specialists habitually read articles back to front, and frequently skip the introduction since they are already familiar with the subject.

**The CARS model.** In 1981 Swales devised a model for the organization of introductions, which he called the create-a-research-space (CARS) model. This model has won general acceptance within the specialty of English for research purposes (ERP), but elements of it have been criticized, prompting Swales to modify the model several times. It was especially reworked in 1990 and 2004. CARS divides the introduction in rhetorical moves that are subdivided into steps. The original 4 moves (Swales 1981) have been reduced to 3, and it has been noted that the moves can go through multiple cycling especially between move 1 and move 2. This revised move-structure has proven usable in many studies of research articles across many different disciplines (Swales 2004). The application of the steps is more varied. Even though already the first model used both obligatory and optional steps, they have not been a perfect fit when used in different disciplines. The current model has incorporated a lot of the stated criticism, and for example included a number of PISF-steps (i.e. Possible In Some Fields – unlikely in others). So whereas the moves seem to be a stable set of descriptors across most if not all fields of science (Swales 2004, p. 226), the steps are of a more local kind. This is why I have chosen to only use the move-structure in this study.

The rhetorical moves are identified by their function in the text, which hopefully corresponds to the intention of the author (Marshall 1987). The rhetorical function is often recognized by grammatical or lexical attributes, or even by their location in the text (Swales 2004, p. 229), but they are not defined by these superficial attributes. Sometimes the rhetorical moves can only be correctly identified by a full understanding of the text. This is why it can be impossible for an outsider to establish the rhetorical structure of a scientific text that requires profound specialty knowledge.

The CARS model is based on studies of English language research articles, and as such it does not necessarily apply to articles in other languages. It does not seem to apply to Polish research articles (Swales 2004, p. 252), and Swales has proposed an OARO-model (Open A Research Option) as an alternative for non-English research.
articles, that may have a ‘softer’ structure due to less competition (Swales 2004, p. 226).

3.7 References and Citations

By references we usually mean formal pointers in writing to existing documents, consistently formed either in parentheses or footnotes. The cited document is said to have received a citation from the citing document (Price 1970).

There are also informal references, references to the collected works of an author or to a research tradition, without noting specific documents (Jacoby 1987; Leydesdorff 1998). Whether such practice is sloppiness or scientific shorthand is anybody’s call.

The obvious purpose of references is to document the sources of the author, and make it possible for the reader to retrieve these sources (Skov 2007). We shall see later that the full picture is rather more complicated.

History. References to earlier works have probably existed almost as long as writing itself, and there are examples of references at least back into antiquity. The footnote also has a long history, in the Middle Ages Scholasticism got its name from scholia, a noting technique dating back to antiquity. At what time the reference and the footnote combined to form the modern reference practice is a debated question (Grafton 1997, p. viii).

Grafton (1997, p. 210-219) sets a decisive shift at Pierre Bayle and Jean Le Clerc in the years 1692-1701. The latter formulated it like this in “Parrhasiana”:

Notes expressed in good terms, in few words, and where one asserts nothing without proving it, or without at least citing some good author where one can see the assertion verified, indicating the passage in question so well that the reader can easily find it, if necessary (cited from Grafton 1997, p. 217-218).

Others postpone the appearance of modern references until the beginning or middle of the 19th Century (e.g. Price 1963), but that might be due to the temporary disappearance of references during the Enlightenment because they were disdained by people like Voltaire who was against these signs of sterile accumulation of data (Grafton 1997, p. 94-95).

The amount and the appearance of references have changed over time. There has also been a development in which kinds of documents references point to. Bazerman (1984, p. 172-175) detects a marked shift in citation practice in the speciality of spectroscopy during the first decade of the 20th Century. Before this time, in the last years of the 19th Century, every article had about 10 references, all of which went to familiar predecessors of the discipline without relating them specifically to the citing article. Furthermore, half of these references were undated. By 1910 the number of references has decreased to 1½ per article on average, but those few citations goes to relevant and current documents. Thus the practice of citing can change fairly quickly, and it seems clear that the new citing practice reflects that the references had attained a new function.
The original purpose of references was to add authority to the citing document. By quoting for example the Bible or one of the church fathers, one hoped to silence disagreement. Later the references acquired a new function, as an information retrieval tool, that could both inform the reader about the existence of certain documents, and perhaps also guide him to find them.

With the advent of the Science Citation Index in the 1960s yet another application was found for references and especially citations. They could now be used for science studies; and reading citations as marks of influence, this can also be utilized for research evaluation. Thus, the use of references has come full circle, where it once was the sources that endowed the citing document with authority, now it is the citing documents that grant recognition to the cited document.

3.8 Citation Theories

With the SCI and the research activity it launched, came a wish for a better theoretical understanding of citations (Kaplan 1965). Until then, there had not been a proper citation theory, but the normative theory of science as described by Robert Merton (1942) and expressed in the norm set CUDOS, implicitly contains an explanation of the occurrence and function of references and citations.

**The Normative Theory of Citing.** This theory originates in the four norms of CUDOS (Merton 1942), i.e. Communism, Universalism, Disinterestedness, and Organized Skepticism. The central norm for the normative theory of citing is communism, meaning that science and knowledge are common property; nobody has rights or privileges over knowledge. The only claim that can be made is the claim of priority. This means that authors should recognize the inventors or discoverers of new knowledge (Kaplan 1965). This can for example be done using eponyms like ‘the Copernican system’ or ‘Boyle’s law’. Merton does not explicitly mention citations, but the principle is the same. Authors should recognize their influences by citing them and in this way pay their intellectual debt.

The other three CUDOS norms are also relevant for citation theory. Universalism, i.e. the requirement to evaluate research regardless of the originator, and disinterestedness, i.e. demand to not have any personal interests involved in ones research, are important elements in citation practice, and as we shall see later, a large part of the critique aimed at citation analysis as an evaluative tool is precisely directed at these norms, as they supposedly are not adhered to in a satisfactorily degree. The fourth norm, organized skepticism, prescribes that scientists test all research rigorously before accepting the results; this applies goes both for their own research and the research of their peers. This requirement can be extended to cover the research practice in general and the citing practice in particular. Seen in this perspective organized skepticism is the guarantee that the other norms will be abided by to a satisfactory degree.

The normative theory of citing sees Science as a normative institution, where the individual citations are recognition and ‘payment’ of intellectual debt; scientists share their knowledge (publish papers) in expectation of receiving reward (citations), e.g.
“The remainder of that debt I shall try to discharge by citation in the pages that follow” (Kuhn 1970, p. xi).

It is obvious that there is a certain discrepancy between the incentive of reward in the form of recognition, and the norm prescribing disinterestedness. This ambiguity is due to the prerequisite that research should be original, and the following importance of priority (Merton 1976, p. 42).

According to the normative theory deviant behavior is very rare, it is expected that almost all authors cite all sources that have influenced them substantially, and that they do not cite documents that have not had influence on the current paper. On the basis of this thinking L. C. Smith (1981) made a list of 5 basic assumptions, which underpin citation analysis.

1) Citation of a document implies use of that document by the citing author.
2) Citation of a document reflects the merit (quality, significance, impact) of that document.
3) Citations are made to the best possible works.
4) A cited document is related in content to the citing document; if two documents are bibliographically coupled, they are related in content; and if two documents are co-cited, they are related in content.
5) All citations are equal.

These five assumptions create a firm foundation on which to use citation analysis, both for research evaluation, information retrieval and science studies. However, the assumptions have been seriously questioned.

**Criticisms.**

The point, however, from the perspective of this paper is that even though citation content analysis is no longer in its pioneering stage, it can claim little more than the discovery that citations are not simple events and cannot be treated as such. (MacRoberts & MacRoberts 1989).

From the middle of the 1970s onwards criticisms against the normative theory of citing have been presented numerous times, and the justification of citation analyses has thereby been questioned. The criticisms concentrate on discrepancies between norms and real-world practice and undermine the credibility of citation analysis. Even to the extent that citation analysis should not be used for research evaluation at all (Wiener 1974; Seglen 1998). Some of the objections are:

**Negative Citations.** It is ostensibly a problem for Smith’s fifth assumption that negative citations do exist, and this criticism has been raised more than once (e.g. MacRoberts & MacRoberts 1989). Several citation studies have tried to measure the amount of negative citations (e.g. Moravcsik & Murugesan 1975; Chubin & Moitra 1975; Oppenheim & Renn 1978; Frost 1979; and see also Small 1982), and the proportion of negative citations in these studies is generally low (1-6%), but in Frost’s and Moravcsik & Murugesan’s studies it reaches 12-14%.

It has been argued that a negative citation can be so obscure that outsiders cannot see it. Gilbert (1977) presents an example of a citation to one of his own works, which he himself deemed to be negative, though the critique is so abstrusely worded, that only readers who are very familiar with the discourse will be able to recognize it.

One would therefore suspect that the proportion of negative citations would increase if authors classified their references themselves. Such a study was performed.
by V. Cano (1989) on the references of 28 research articles. The authors of the articles classified only 2% of the references as being negative. If the negative citations are randomly distributed between all mediocre or worse papers then 2% negative citations would make no difference in the statistics. But if the same document get cited negatively repeatedly, its citation rate would be boosted disproportionally. That this perhaps is the case is indicated by the so-called J-curve (Nicolaisen 2002), which shows that inferior documents have higher impact than mediocre documents.

**Padding.** Padding is the practice of adding superfluous references (Cronin 1981), there can be various motivations for this. The author may hope to strengthen the credibility of the paper by having a long reference list including authoritative works (LaTour 1987). The author can add excessive citations to himself, or he can cite others for social reasons. The author may also add citations to a specific journal in order to get his paper published there. There might also be purely ceremonial citations.

In studies of the typology of citations (Small 1982) perfunctory citations range from 20-60%, though it should be noted that the 60% was reached in a study of citations to old physical science papers, and most of the perfunctory citations were for historical reasons (Oppenheim & Renn 1978). The studies (Moravcsik & Murugesan 1975; Chubin & Moitra 1975; Oppenheim & Renn 1978; Frost 1979) do not use the same classification system, which makes them only partly comparable. Still it seems that between one and two fifths of citations are perfunctory. This is quite a high rate, and shows the potential value of a method that could pinpoint the most relevant references in a text.

**Flattery.** It is conceivable that authors, who wish to make a good impression for example on editors of high-end periodicals, could equip their articles with redundant references to them. This has been claimed by P. O. Seglen (1998), but a methodical study by Faber Frandsen & Nicolaisen (2011) has failed to support that hypothesis.

**Self-citations.** It has generally been seen as a problem that authors cite themselves (e.g. Seglen 1998). However, self-citations are not necessarily signs of bad form, as White (2001) points out. The author’s own works are well-known to him, and thus make them convenient to reference. Furthermore, as authors specialize in specific subjects, often the most relevant previous works will be the author’s own work.

All the same, self-citations are still viewed with suspicion by some agents of academia, for example do the guidelines for submitters to the Journal of the American Chemical Society specifically warn against excessive self-citations (American Chemical Society 2011).

**Omissions.** It has been pointed out repeatedly that authors do not cite all their influences, this can among other causes be due to ignorance of the literature. MacRoberts & MacRoberts (1986) argued this claim in a study of 15 articles in genetics. They found that only 216 out of 719 necessary citations were actually made. Judging which and how many references are ‘necessary’ to cover a phrase seems to be a very subjective process. The MacRobertses give some examples on how they made these decisions:

> Here we cite two reviews covering and supplanting fifty or so papers, many of which we had read and were influenced by. We do not cite them because it is more parsimonious to use the two reviews. … In these cases, what does the analyst do? If
he is concerned with influence, then references must be added to bibliographies (MacRoberts & MacRoberts 1986, p. 155).

And:

Weinstein is one of the few Mendel scholars publishing in the English language who persists in going to original sources. This is an admirable habit and should be commended. But for the analyst it presents problems because it leaves out intermediate sources of influence or use. For example, in Weinstein's paper mentioned above, we find no explanation of how he located the esoteric work of Schmalhausen, Besnard, Blomberg, and others. It is unlikely he did so by a diligent search of the nineteenth century literature; instead he probably used uncited intermediate sources. These intermediate influences are omitted from his bibliography, and the analyst presumably must reinstate them (MacRoberts & MacRoberts 1986, p. 161).

These examples suggest that the MacRoberts’s list of ‘necessary’ references was very comprehensive, and perhaps exaggerate the problem.

Obliteration by Incorporation. Obliteration is the phenomenon that a concept that is well-known within a field is mentioned without a formal reference to the source document because it is expected that all subject specialists are familiar with this document (Swales 1986, p. 39; White 2004, p. 104). The concept has been incorporated into the background knowledge of the field (Hyland 2000, p. 20).

Many of these incorporated concepts are now referred to eponymously after the creator or discoverer; examples include Bradford’s law, Pearson’s r and the Planck constant. In these cases the source documents can be recovered fairly easily, but other non-eponymous obliterations are more obscure, such as the Matthew effect (Robert Merton), Web 2.0 (Tim O’Reilly) and Family resemblance (Ludwig Wittgenstein), it can be difficult to track down the original source for such concepts. Needless to say, obliteration has a detrimental effect on citation counts.

Bias. MacRoberts & MacRoberts (1989) claim that there is a clear tendency that some concepts get credited almost every time they are mentioned, while others are repeatedly neglected, or is only cited through a secondary source. The MacRobertses do not explain the phenomenon themselves, but it must be related with obliteration and citations as concept symbols (Small 1978). Apparently conventions guide how concepts are referred to, and the conventions vary for different concepts. Some concepts are referred to eponymously, others by a distinctive term, and yet others by a customary citation. When the citation goes to a secondary source, the reason can be that the original source is difficult to access.

The reasons why concepts are referred to in different ways and what guides which method become customary for specific concepts in specific discourse communities has not to my knowledge been thoroughly studied, but there certainly is a citation bonus for the concepts that are customarily invoked with a citation.

Citations as Persuasion. The abovementioned criticisms weaken the normative theory, and have led to an alternative citation theory, the idea that references are primarily included in order to persuade the reader of the validity of the claims in the citing paper (Gilbert 1977).

Gilbert reads references as an inherent part of the citing text. References are intentionally included by the author, and as such they serve the same overall purpose
as the work they are situated in. According to Gilbert, the purpose of research articles is to persuade the scientific community of the reality of the claims in the article. This follows from the main purpose of research, which is to contribute to the collected scientific knowledge.

To become incorporated in the collective scientific knowledge, a claim must be accepted as original, important and true. In order to achieve this, the author locates his paper in relation to current knowledge to show that the current paper fills an important knowledge gap, and then presents evidence to persuade the reader of the reality of the claim. References are the standard way of presenting current knowledge and indicating the gap that one wishes to fill, but references are also an efficient tool in the persuasion of the reader. By referring to already acknowledged works, the author lends support for his claims and gains more credibility.

References are used to argue the currency, the importance and the credibility of research. The author uses the references to substantiate the individual claims of the paper. Furthermore Gilbert (1977) claims that the aggregated references in the form of the reference list are used to boost the persuasiveness of the article.

In order to be persuasive it is important to cite the right documents in the right way, the author must cite in accordance with the predilections of the readers. If the author cites an authoritative source to substantiate a claim, it is only persuasive if the readers also regard that work as authoritative. In effect, the author shows which discourse community he belongs to by his references, this makes his references more persuasive for readers adhering to the same discourse community, and reciprocally the citations strengthens the discourse community. In other words, it is not simply the individual author who chooses which works is important and should be cited, it is the discourse communities and by extension the scientific community (Cozzens 1981; MacRoberts & MacRoberts 1986; Hyland 2000).

Different references have different connotations for different readers. The choice of references is done on the basis of the author’s expectation that the readers will regard them as credible and relevant. This expectation can never be more than an educated guess, as the author does not control who reads his article (Gilbert 1977).

Luukkonen (1997) argues the validity of the persuasion theory by presenting three observations which contrary to other theories is explainable by the persuasion theory:

1. The heterogeneous usage of references. In spite of the fact that authors have different motives and different rhetorical reasons for inserting references into their texts, and these reasons vary from one reference to another in the same text, the general role of citations is uniform, that of supporting knowledge claims. The heterogeneity is indicative of the variety of means used to fulfil this function.

2. The difference between the intentions and interests of the authors of the cited and citing texts is being explained - knowledge construction is a social process and authors of texts cannot control the later usage of their texts. The subsequent transformation of the cited texts in the process is an essential part of the social construction of scientific knowledge.

3. Different citation 'etiquettes' are understandable considering the primary function of citations as a rhetorical device and the fact that different groups of scientists might have different discourse practices.
Luukkonen presents these three points in favor of the citation theory put forth by Bruno LaTour (1987), but the points seem to be equally applicable to the propositions of Gilbert (1977).

In LaTour’s theory of the sociology of science, the function of references is persuasion, thus it fundamentally agrees with Gilbert. But LaTour takes the idea of persuasion to the next level; references are not indication of sources, but ammunition in the persuasion of the reader (White 2004) and what becomes established as knowledge has more to do with discourse than any link with reality.

Padding is a generic trait of scientific papers. References are needed to show that a document is serious; the more references the greater a threat they pose to a potential dissenter, bringing strength in numbers. Therefore many citations are perfunctory; they are only in the paper for display, or to mark which discourse community the author belongs to.

References are not just arrayed, they are also modalised; they are put into a context that qualifies the cited work, and in this context “all deformations are fair” (LaTour 1987, p. 40). Not only can a cited work be criticized, it can also be misquoted. The amount of superfluous and misquoted citations means that if a critical reader performs the overwhelming work of tracking all the references in an article, the result could be disastrous for the author (p. 33).

However, the worst that can happen to a scientific paper is to be ignored, the value of the article is decided by the users, and if there are no users the value is nil. LaTour claims that the majority of papers have no impact whatsoever as they are never read (LaTour 1987, p. 40; see also Paul 2000).

If an article does escape the fate of being ignored, it will be used and abused as reference in later papers to serve their purpose, and it is in this later use and the possible transformation that happens at this stage, that the purport of the article is decided by the discourse community. The cited and the citing papers mutually strengthen and modify each other, and this process results in the creation of black boxes: undisputed facts. Fact construction is a social process (LaTour 1987, p. 41). The discourse community is thus the central element in science, taking precedence over research.

**The Mantra of Average.** The criticisms coming from social constructivists can be mitigated by looking at citations in the aggregate. Instead of studying what the intentions are for each individual citation, one might look at what the combined effects of the totality of citations indicate. This view can be supported by an analogy: You do not have to follow every single molecule in order to describe and predict how a quantity of gas behaves under different conditions (Price 1963, preface).

This approach can also be illustrated by this quote:

> The issue is not whether we can rely on reference lists in individual cases as complete sets of influences (we cannot), but rather whether references can be used statistically, in the aggregate, as an indicator of influence. (Small 1987)

It is possible to find many examples of manipulation and sloth in citation practice: excessive self-citations, copy-pasting of references, superfluous references, omissions, etc. But when we aggregate enough citations these elements of error will even out supposing that they are randomly distributed. So we can still do citation analyses, and use the results for research evaluation.
The important thing to note here is that there will be a considerable amount of background noise in citation analyses, and to get reliable results from the data, we must be cautious when gathering the data and performing the calculations, we must control the variables we can and be certain that uncontrollable variables are randomly distributed before determining if the result is significant.

The mantra of average does not have a theoretical basis, except for a maxim that the general reason for citing is that the citer considers the cited document to be relevant in the context of the textual passage that holds the citation (e.g. Liu 1993, p. 379; White 2001, p. 98; Swales 2004, p. 85). Succinctly formulated by White (2004, p. 107) in this way:

The global reason for citing is the perceived relevance of a document to a work in hand.

Citations as Concept Symbols. According to Small (1978) citations must be understood in the context of the location of the citation in the citing document. References acquire meaning in the light of the word or phrase they are placed next to. The author uses the reference to impart additional information to his text, at the same time he endows the cited document with additional meaning, since the citation in itself holds information about the cited document. The author might state his opinion of the cited document, or set it in a new perspective, and even if the citation is uncommented, it still holds the information that the author considers the cited document to be relevant in connection with the citing text.

There is a certain normative control on this process. There might be other motivational factors for the author to cite certain documents (flattery, padding etc.), but as the reader expects a cognitive link between the citing and the cited text, the author is not at liberty to cite any document he choose, he must ensure that the citation makes sense at that specific point in the text.

The main point of Small’s theory is that authors have a register of documents that they are very familiar with, and that the concepts an author commonly use will be linked to these documents in his mind. This means that when the scientist wishes to evoke a certain concept in an article, he will routinely cite the document that in his mind is linked to that concept; the document has become a concept symbol for the author (Small 1978). Such a register facilitates referencing, as it makes it easier for authors to find a relevant document for the text passage in question.

Inspired by Relevance Theory H. D. White see relevance as cognitive effects combined with ease of processing, in this light the abovementioned author chooses the document that is most relevant to him in the context as the ease of processing is much higher for an already known document (White 2007). What would be the most relevant document in the context for the reader is thus not necessarily the same as the most relevant for the author. For example will the author’s own production have very high ease of processing and therefore high relevance for the author, though this might not be the case for the reader.

When a reference is used as a concept symbol it strengthens the link between the cited document and the concept, the author is more likely to reuse a concept symbol the more he has used it already, and the reader will also be more likely to link the document and the concept the more he has seen them paired in a paper. If more authors use the same concept symbol it changes status from a private to a common
symbol, and can gradually become a standard within the discourse community. When citations acquire this status they resemble descriptors, a resemblance that has been employed by ISI to create the ISI’s KeyWord Plus (Garfield 1990b).

When it becomes a standard in a discourse community to link a certain concept with a certain document an author can evoke the concept for the reader simply by citing the proper document. This can be used to shorten the text and keep focus on the line of thought without unnecessary digressions, though this easing of communication does happen at the expense of simplifying the contents of the source document, which is reduced from a complicated rhetorical act to an unambiguous symbol for an established scientific concept. It is even possible that the symbolic significance a document holds for a discourse community is incongruent with the message the author of that document tried to communicate, see for example Cozzens’ (1981) description of the concept symbol of (Kaplan 1965) in citation theory.

This phenomenon that a citation becomes a symbol for a certain concept is comparable with obliteration by incorporation; it is simply another way to evoke a complicated concept with a word or a short phrase. The classic example of a citation used as a concept symbol is Lowry’s article on protein measurement (Lowry et al. 1951; Garfield 1990a), which is quoted by almost 300,000 articles in SCI, and still receives more than 1000 citations per year. All these citing articles do not discuss Lowry’s paper, most simply write ‘(Lowry et al. 1951)’ as a concept symbol.

Citations as concept symbols are usually limited to certain discourse communities. Therefore it is possible to identify discourse communities by analyzing how different authors use citations as concept symbols. For example, co-citation analyses have proven to be an efficient tool to create overviews of scientific disciplines.

**Historical Afterthought.** The citation theories were only formulated in the 20th Century, but many of the notions behind them have been around for centuries, as the following three examples illustrate.

The existence of a norm dictating recognition of precursors is demonstrated by Samuel-Auguste Tissot in 1768. He had in discordance with the fashion of the time supplied his text with numerous footnotes, and he justified this practice by stating that he inserted the footnotes to give the authors whose works he had used ‘the honor he owed them, with some words set at the margin, where they hurt nobody in any way’ (Grafton 1997, p. 95-96).

The persuasion theory is foreshadowed in this quote: ‘I felt citation was indispensable in the work of a beginner who has to make his way and earn confidence’ Leopold von Ranke (1795 – 1886) according to Grafton (1997, p. 64).

Also the many criticisms raised by social constructivists have precursors. In 1778 H. E. Davis writes a long disapproving review of Gibbons’ ‘The Decline and Fall of the Roman Empire’. Davis claims that Gibbons distorts references so that they seem to support his view where in fact they do not; that he trims quotes to avoid unwelcome facts; that he does not reference secondary sources that he had in fact used; and that he plagiarized. According to Davis, Gibbons’ primary use of footnotes was to avoid being exposed.

**The Need for a Unified Theory of Citation.** The call for a theory of citation has been reiterated several times (e.g. Cronin 1981, Nicolaisen 2007, p. 611). The question is whether an established citation theory, even if it could be agreed upon which seems unlikely, would be an advantage. For example, John Swales made a
clear definition of genre in 1990, but later (2004) he realized that such a definition, even though it was a good and functional definition, results in blind angles. This is why he substituted this definition with a number of loose metaphors (Swales 2004). In a like manner it might give blind angles if we ever reach an agreement of a unified citation theory.

Obviously, on a very basic level we do know exactly what references and citations are: A reference is a link pointing from a location in one document to another document, and a citation is the reciprocal mentioning of a document in a later document.

I will in the following section try to analyze references and citations in research articles by combining this basic definition with what we know of the research article genre.

3.9 Classification of Citations Based on Genre

It seems to be a profitable approach to see citations and references as elements of the citing text. References are intentionally included by the author and must therefore be assumed to be a part of the rhetoric of the text, and their function to be to further the purpose of the document. To explore this further, I will in this section look at references and citations from the perspective of the genre of the citing document.

It is paradoxical that even though genre is not generally an element of citation theories, it is implicitly included because most papers on citation theory almost exclusively use one genre namely research articles as examples. Likewise, with few exceptions (e.g. Frost 1979), the objects of citation classification studies are research articles. A reason for this bias might be that citation analyses are usually performed in SCI, in which 72% of the documents are classed as articles.

Another reason that might bear more weight is that the research article is the scientific document par excellence of today, it is used in all fields of science, and in most fields it is among the most important formal genres. The research article is placed in the very top of the hierarchy of scientific document types, and this might explain why so many authors take it for granted that when we talk citations we mean citations in research articles.

References in Research Articles. When we look at references from the vantage point of the genre of research articles, it does offer some immediate insights. The first rather banal realization is that the primary function of references in research articles is to indicate the scholarliness of the article (Skov 2007). References are compulsory for research articles. A letter or an editorial can get published without references, but for articles references are required by tradition.

Secondly, references are an element of the text and as such its function is to further the purpose of the article. This is not specific for research articles as it is the case for all genres that use references. But it does mean that we must look closer at what the purpose of research articles is.

Besides being the most important formal channel of information for large parts of the scientific community, research articles also play a crucial role in the codification process from research to scientific knowledge. Scientific discoveries do not hold any
value until they are shared with the scientific community (Merton 1942, p. 274). Research must be published, so that it can be tried by the scientific community and perhaps become part of the collected scientific knowledge (Small 1978, p. 338), though it is also quite possible that the research never comes further than publication, as most articles never receive any citations.

If we presume the primary purpose of science is to contribute to the collected scientific knowledge, the purpose of research articles is to present knowledge claims in a persuasive way. That is, to convincingly argue that the presented research is important, original and true. References are used to support the discourse in order to achieve this purpose.

**Functions of Citations and References.** The above reflections lead to three distinct functions for references and citations in research articles.

Citations place the citing document in relation to current knowledge. Research articles conventionally begin with a review of earlier publications in the same research area, by referring to these documents the author shows what the previous research has not done yet, and thereby advocate the originality of the research at hand.

Citations are used communicatively. The rhetoric of scientific papers must be unambiguous, precise and concise. The author has only a limited number of pages available and references can be a helpful tool here. Scientific knowledge that is already commonly accepted, what LaTour calls a *black box*, can be imported in the rhetoric with a brief “it has been shown that ... (NN 2000)”. The author does not have to present independent evidence for the claim, saves room to argue the central points of his article, and avoid a digression that could impede the reading of the text.

Citations are used persuasively to help achieve the overall purpose of the article, which is to convince the reader that the research is original, important and true – or at least: new, interesting and probable. According to Gilbert (1977) “authors will tend whenever possible to cite papers which they consider their audience will regard as presenting valid and important arguments and results.” An author will primarily cite works which he assumes that the reader will regard as authoritative. Since readers, authors and reviewers in this world to a great extent are the same people in different combinations, a discourse community will develop a common understanding of which documents are authoritative and what concepts they apply to (Cozzens 1981), this is why a discourse community will often use the same citations as symbols for the same concepts (Gilbert 1977).

A fourth function of citations in research articles is the aggregated references in the reference list. According to Gilbert (1977) the content of the reference list can be used to catch the attention of potential readers as many scientists read the reference list first in order to judge if it is worth the effort to read the article proper. The reference list might also be used to intimidate potential critics (LaTour 1987), or to flatter journal editors and thus make the article more publishable. However, it should be noted that none of these claims have been substantiated by empirical research.

**Codifying Other Works.** The abovementioned functions for citations in research articles are seen through the author’s wish to establish his article in the collected scientific knowledge, but the research article has a dual role in this process; not only are articles trying to win a place for themselves, at the same time they are establishing other documents as scientific knowledge. And this is done through citations.
Citations emphasize the documents they cite. The readers of the citing document are now aware of the cited document, and the cited document has earned one more citation, which makes it slightly more visible in citation indexes. Thus the cited document is slightly more established than it was before the citation, whether that was the intention of the citing author or not.

This process can be seen as repaying of intellectual debt or as black boxing, depending on perspective.

**Justification of Citation Analysis.** In this perspective citation analyses is a rational tool for science studies and research evaluation. Citations will mainly go to ‘good’ documents, i.e. documents that are recognized by the discourse community, or documents that the citing author expects that the discourse community will find reliable (Hyland 2000). Though some works get cited so that the author can criticize them, it is reasonable to expect that the majority of citations go to documents that can be used to support the current article.

Furthermore, since citations are a major way to establish what is accepted as facts by a discourse community, citing creditable documents is a self-reinforcing action as the cited documents become more creditable by being cited.

Citations also indicate a probability of an intellectual link between the citing and the cited document. Both because the author is likely to cite works from his own discourse community as per above, but also because citations are brought in as evidence for specific claims, and a cited document that does not substantiate the claim in question, will be a liability to the author. In the words of Bruno LaTour:

> It [stacking references] might be a source of weakness. If you explicitly point out the papers you attach yourself to, it is then possible for the reader … to trace each reference and to probe its degree of attachment to your claim. And if the reader is courageous enough, the result may be disastrous for the author. (LaTour 1987, p. 33)

This is not to say that all citations indicate quality and relevance, but it is reasonable to expect that a substantial part of citations does reflect what is considered by the appropriate discourse community to be creditable and relevant in the context. Which is to say that citations are not indicative of absolute quality, but are reflections of the values of the discourse community. This is why there is correlation between citations and other signs of academic recognition (Cole & Cole 1973).

With this theoretical basis we can use citation analysis to identify semantic relations between documents and to identify documents that are held in esteem by certain discourse communities, though we cannot make the reverse conclusion that lack of citations means no scientific influence.

Some citations are problematic, authors may distort citations in order to increase the persuasiveness of the article (LaTour 1987, p. 40), and other motivational factors can also influence the author when he adds references: self-promotion, padding, flattery and so on. But tendencies in these directions are tempered, both by the gatekeepers - reviewers and journal editors, and by the reader’s expectation that references will be relevant (Small 1978, p. 328). If the reader finds that citations are redundant, irrelevant or distorted, it will decrease the credibility of the citing author.

The readers and gate-keepers form a control measure on the citation practice in journal articles. The control measure does not prevent all ‘cheating’ but does ensure
that a reasonable proportion of citations is in accordance with the norms (Nicolaisen 2004, p. 164).

4 Results and Discussion

To avoid that the results section become unnecessary verbose, I will in the following section use the term ‘reference’ to signify the document that the reference points to, likewise the word similarity is always referring to the similarity between one of the references and the source article citing it, measured either as the vector based similarity score or the tiered vector based similarity score. For example the sentence “the references in the introductions show more similarity than the references in the method sections” should be read as “the documents cited in the introductions have higher vector based similarity scores with the citing source article than the documents cited in the method sections”.

4.1 Description of Source Articles

The data in this study is derived from the 30 most cited articles in volume 122 of Journal of the American Chemical Society. The reason to choose the source articles by citation count is that poorly cited articles would provide poor data, for example would an article without citations not add any information to the study. The reverse is also true, highly cited articles have larger potential for high co-citation counts, which will provide more solid data. The source articles have received from 224 to 937 citations in SCI.

The articles generally follow the IMRAD structure: Introduction, Method, Results And Discussion. All articles begin with an abstract and end with ‘Acknowledgment’ and often also a short section with a link to more comprehensive research data on the internet. The method section is usually called ‘Experimental Section’. In seven of the articles it is placed at the end of the article sometimes in a smaller font, as an indication that method sections are regarded as less important.

10 of the articles have a results section and a discussion section, but the rest have a combined results & discussion section. Therefore I do not discern between the two sections but aggregate them. 20 of the articles end with a section called ‘Conclusion’ or ‘Summary’, only seven references are located in these optional final sections, so I have chosen not to analyze them independently. Therefore the references from conclusions and summaries have been added with six other incongruent references: four from appendices, one in the article title and one orphan, i.e. there is no footnote number in the text indicating where the reference belongs.

References are placed numerically under the text body. Apart from that, the reference practice varies widely. 8 articles use the footnotes exclusively for references and have one footnote per reference, but it is common practice to place multiple references in one footnote, and using the footnotes for explanations, references to other footnotes or web-links. In total the 30 source articles have 1399 footnotes containing 1723 references and 6 web-links.
Roald Erik Frosg

Citation Classification Based on Genre

Table 2. Characteristics of the 30 source articles, all published 2000 in J. Am. Chem. Soc.

<table>
<thead>
<tr>
<th>Article#</th>
<th>First Author</th>
<th>No. of authors</th>
<th>Citations in SCI</th>
<th>References</th>
<th>Number of highly cited references</th>
<th>move-structure based on paragraphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Manna, E. L.</td>
<td>3</td>
<td>937</td>
<td>36</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>Mattoussi, H.</td>
<td>7</td>
<td>783</td>
<td>52</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Litke, A. F.</td>
<td>3</td>
<td>824</td>
<td>96</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>Storhoff, J. J.</td>
<td>6</td>
<td>563</td>
<td>42</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>Rumi, M.</td>
<td>13</td>
<td>474</td>
<td>49</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Ediaoudi, M.</td>
<td>3</td>
<td>493</td>
<td>28</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>He, L.</td>
<td>7</td>
<td>477</td>
<td>59</td>
<td>27</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>Kepert, C. J.</td>
<td>3</td>
<td>355</td>
<td>60</td>
<td>39</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Lu, Y. F.</td>
<td>7</td>
<td>336</td>
<td>27</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>Ruckert, M.</td>
<td>2</td>
<td>314</td>
<td>25</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>Strether, T.</td>
<td>5</td>
<td>295</td>
<td>44</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>Brinkmann, M.</td>
<td>6</td>
<td>280</td>
<td>68</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>Beauvais, L. G.</td>
<td>3</td>
<td>284</td>
<td>82</td>
<td>40</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>Eckert, J. F.</td>
<td>10</td>
<td>271</td>
<td>93</td>
<td>46</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>Imperor-Clerc, M.</td>
<td>3</td>
<td>280</td>
<td>40</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>Grönbeck, H.</td>
<td>3</td>
<td>282</td>
<td>32</td>
<td>27</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>Seminaro, J. M.</td>
<td>3</td>
<td>270</td>
<td>40</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>Min, K. S.</td>
<td>2</td>
<td>283</td>
<td>28</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>Dembinski, R.</td>
<td>5</td>
<td>266</td>
<td>96</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>Vargas, R.</td>
<td>4</td>
<td>260</td>
<td>92</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>21</td>
<td>Tsuzuki, S.</td>
<td>5</td>
<td>260</td>
<td>81</td>
<td>38</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>Guckian, K. M.</td>
<td>6</td>
<td>257</td>
<td>64</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>23</td>
<td>Benoi, D.</td>
<td>6</td>
<td>256</td>
<td>72</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>24</td>
<td>Tempel, D. J.</td>
<td>5</td>
<td>250</td>
<td>80</td>
<td>35</td>
<td>23</td>
</tr>
<tr>
<td>25</td>
<td>LaBean, T. H.</td>
<td>7</td>
<td>225</td>
<td>46</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>26</td>
<td>Struijk, C. W.</td>
<td>13</td>
<td>340</td>
<td>82</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>27</td>
<td>Tsui, V.</td>
<td>2</td>
<td>236</td>
<td>56</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>28</td>
<td>Zhang, Q.</td>
<td>3</td>
<td>242</td>
<td>52</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>29</td>
<td>Lynn, D. M.</td>
<td>2</td>
<td>237</td>
<td>49</td>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>30</td>
<td>Mikulec, F. V.</td>
<td>6</td>
<td>224</td>
<td>51</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>153</td>
<td>10754</td>
<td>1722</td>
<td>849</td>
<td>478</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>5.1</td>
<td>358.5</td>
<td>57.4</td>
<td>28.3</td>
<td>15.9</td>
</tr>
</tbody>
</table>

The 30 source articles differ in how many references they contain and how the references are distributed over the sections. Also, the articles are written by different authors having different writing styles, which can affect the data. For example, there is a marked difference in sentence length (see appendix B), which has an effect when coding moves on sentence level. I have not marked this difference of writing style, but assume that the variance it results in is randomly distributed, and that the articles in the aggregate give a valid picture of the total population.

**Atypical Articles.** Five of the articles deviate from the conventional structure. I will here describe them and state how I have coded these discrepancies in SPSS. The bibliographic information on the articles can be found in appendix A (p. 54).
Article#3 has a rather idiosyncratic structure, it is arranged in three sections: ‘Introduction’, ‘Results and Discussion’ and ‘Conclusions’. The method seems to be incorporated in the ‘Results and Discussion’ part, furthermore, the two first paragraphs of this section (§4 and §5) are devoted to background information that seems to belong in the introduction. I have therefore chosen to code §§4-5 as part of the introduction when assigning moves on the levels of sentence and paragraph, though not in the algorithmic coding.

Article#9 is only 4 pages long and is not divided into sections, but content-wise it does follow the IMRAD structure, so I have coded it into sections accordingly (See Appendix B, p. 62).

Article#15 has a theory section called ‘Structural Models’ between the results section and the discussion section. I have chosen to include this section in the results & discussion section.

Article#17 is a theoretical study; I have chosen to code the two sections ‘Theoretical Approach’ and ‘Background Studies’ as method, and ‘Results on the MRTD System’ as results & discussion section.

Article#27 has a mathematical section called ‘The Generalized Born (GB) Model’ which I have coded as method. It is also the only source article to have references in the appendix.

Errors in SCI. As I gathered the data from SCI via Dialog I encountered some errors and omissions, from relatively harmless mistakes in page numbers to missing references and spelling errors in author names.

The 30 source articles contain a total of 1723 references, but in SCI they have only 1699. 26 references are omitted in SCI, on the other hand there are two mistaken additions: In article#19 the reference to ‘Quiros Mendez, N.; Seyler, J. W.; Arif, A. M.; Gladysz, J. A. J. Am. Chem. Soc. 1993, 115, 2323’ has been transformed to two references in SCI under the author names of MENDEZ NQ and GLADYSZ JA respectively. The other addition is from article#24, which includes two references to the same patent. Therefore, the empirical base of this study ends up being 1697 references.

I did not methodically look for errors in the dataset, but I came across a total of 9 typographical errors, e.g. Maggini becomes MIGGINI and Azis becomes ARIZ. The distribution of errors is: 4 spelling errors in author names, 2 listed under second author instead of first, 1 spelling error in title, and 2 wrong page numbers. It should be noted that for one of the four discrepancies in author name it is likely that the reference list in the article has it wrong and SCI has corrected it: Article#21 has a reference to L. A. Curtiess, a name that does not occur in SCI, neither as author nor as cited author, SCI list the reference as CURTISS LA, and the article has 486 citations in SCI under that name, making it likely that SCI has the name correct.

The 8 other references with errors has from 1 to 16 citations in SCI under the erroneous entry, I did not check how many citations were attributed to the real name.

To sum up: 34 references are missing or erroneous out of 1722, which amount to an error rate of 2%, which is an acceptable proportion of error, since the errors can safely be supposed to be randomly distributed.

Variables. In SPSS I have attributed the following variables to the references: location in the text, age, number of citations and number of co-citations. The location
in the text is coded in four variables: section, move coded on sentence level, move coded on paragraph level, and move algorithmically coded. I have chosen not to include the document type of the reference, as that would have been very time-consuming. The references go to a variety of document types including articles, patents, monographs and conference papers. There is a marked difference in the amount of citations these different document types accrue, and probably the meaning of co-citations in relation to citations will also vary across types. For example, a co-citation from a monograph is more likely to stem from two different and unrelated sections than a co-citation from a short article. Hopefully, references to different document types will be distributed randomly across the moves and sections of the source articles.

4.2 Overview of the Empirical Data

References. The number of references in the source articles varies from 25 to 96, median 57 (Table 2, p. 42). The majority of references are placed in the introduction where we find 903 out of 1697 which is 53%. Method sections contain 205 references (12%), and results & discussion sections 34%. The remaining 1% are located in conclusions and appendices. The oldest reference goes back to 1904; the median age is 1996 and the average 1992.

![Distribution of citedness of references. Frequencies are given on a logarithmic scale.](image)

**Fig. 3.** Distribution of citedness of references. Frequencies are given on a logarithmic scale.
Citations to References. All references have received a minimum of one citation, as all references are at least cited by the source article. The highest-cited reference has 31,125 citations in SCI, the average number of citations is 499, and the median is at 81. I have chosen to use the median to separate the low-cited references from the highly-cited, so that highly-cited in this study means 81+ citations. The analysis focuses on the highly-cited references, as random deviations have much greater impact on low citation counts than on high counts. The number of co-citations shared by source articles and references ranges from 0 to 500 averaging 16 and with the median at 4.

As we see from the histograms (Fig. 3 and Fig. 4) both the citedness and the co-citedness distributions are exponential. This makes the statistical analysis less reliable as most statistical tools work from the assumption that the distribution follows a bell curve. Therefore the highly-cited references have been tiered by similarity score in 10 groups of equal size.

![Fig. 4. Distribution of co-citedness between references and source articles. Frequencies are given on a logarithmic scale.](image)

Vector Based Similarity. As described earlier, co-citation counts are weighted to compensate for received citations by dividing the co-citation count by the square root of the product of the citation counts of the two co-cited documents, i.e. the source article and the reference (see formula on page 16). The following analysis is limited to the highest-cited half of the references, the vector based similarity for these references go from 0 (no co-citations) to 0.47 (474 and 1056 citations respectively and 330 co-
citations). The similarity variable correlates significantly ($p < 0.05$) with the year of the references (Table 4, p. 81). The correlation is small to medium at 0.297, but even so I include the reference age in the following analysis, otherwise I would not be able to tell if for example the references in the method section are less co-cited with the source article solely because they are older than the average reference or if the location has a predictive effect independent of the age of the reference.

First I examine the distribution across sections of the references’ similarity scores. Then I will take a closer look at the introductions, examining the variations of similarity over the three moves. This part of the analysis is done three times, once for each technique by which I have determined the moves. So I am examining the distribution of similarity scores across reference location in four stages: by section, by move determined on sentence-level, by move determined on paragraph-level, and by move determined algorithmically.

For all four stages I analyze the distributions across location – either sections or moves - with a series of graphs and tables, which can be found in Appendix C on page 81. The distributions are first illustrated with a boxplot of vector similarity scores across locations, and descriptive data are provided for age and similarity. This is followed by a histogram of tiered similarity groups and a contingency table with tiered similarity distributed over locations and separated in old and new references. Finally a $\chi^2$ test is presented to show whether the locations have a significant predictive influence on the vector based similarity, this test is divided in old (published 1904-1995) and new (published 1996-2000) references.

### 4.3 References Across Sections

Fig. 5 shows the correlation between co-citations and citations. The slanted line marks the perfect cases where all citations to the reference also are co-citations to the source article, the closer the reference lies to this line; the greater is the proportion of co-citations to citations.

We can see that though there is a large overlap, the references in the three main sections of research articles have markedly different trends. The method-references lie more to the right and downwards, signifying that they are more cited (cf. Small 1978, p. 334), but also proportionally less co-cited. In the method section we also find the oldest references (Table 5, p. 82) which corresponds with White’s (2004, p. 104) statement that methodical papers get the highest and the longest-lasting citation counts. The references in the introductions are closer to the perfect line, they have slightly lower citation counts, and the highest co-citation counts. The references in results & discussion seem to be more downwards to the left in the picture, so they are less cited, but their similarity to the citing document will be easier to establish in the other graphs and tables.

The boxplot graph (Fig. 6, p. 81) shows that the vector based similarity is distributed exponentially; that can be seen by the low placing of the medians and the many outliers and extreme values in the top range. The difference between the sections is evident in this graph; introductions have markedly higher and the method sections lower similarity scores, the results & discussion section lying in-between.
Introductions are the most reference-dense section in research articles, containing more than half of the total number of references. The references in the introduction are the most recent, and score better on the vector based similarity, averaging twice as good as the other sections (Table 5, p. 82). The average similarity score for highly-cited references in the introduction is 0.094, for the method section the corresponding score is 0.029 and for the results & discussion section it is 0.051.

![Co-citations compared with citation count of the reference, shown on a double logarithmic scale, and color-coded according to section. The ideal line (co-citations = citations) is added.](image)

**Fig. 5.** Co-citations compared with citation count of the reference, shown on a double logarithmic scale, and color-coded according to section. The ideal line (co-citations = citations) is added.

When we look at how the sections are spread across the tiered similarity groups (Fig. 7, p. 83), it is clear to see that the highest scoring references are generally cited in the introductions, and that the opposite is true for the method section. The median tiered similarity for introductions is tier 7, for method and results & discussion the median tiers are 3 and 4 respectively. The $\chi^2$ test (Table 7, p. 84) shows that the results are significant ($p < 0.05$) both for old and new references.

Thus, it is clearly the references in the introduction that are most interesting to study closer, both because the introductions hold the most references numerically, but primarily because it is in these sections we find the references with the highest similarity scores.

### 4.4 References across Moves

**Coded by Sentence.** The 30 introductions were read through, and every sentence ascribed to a move, thereafter the references were coded with move corresponding to
the sentence they are located in. The detailed assignment of moves on sentence-level can be seen in appendix B on page 56.

When the moves are determined on sentence-level the cycling nature of the move-structure becomes very apparent. As mentioned on page 13 most references become coded as move 1 when using this technique, and we see that the distribution of the 500 highly-cited references in the introduction is: 422 in move 1, 63 in move 2 and a mere 15 in move 3 (Table 8, p. 85). Move 2 references have the lowest similarity scores, and move 1 references show slightly more similarity than move 3. However the skewness of the distribution means that random variations have too much influence and as the $\chi^2$ tests show (Table 10, p. 87), the results are not significant, so there is no reason to examine these data further.

**Coded by Paragraph.** For this section the moves were determined paragraph by paragraph, and then the references were coded according to the paragraph they are placed in. This gives a better distribution of the highly-cited references across the moves: 241 in move 1, 177 in move 2 and 82 in move 3 (Table 11, p. 88). On the boxplot (Fig. 10, p. 87) we see that move 3 have the highest median on the similarity scale, but both move 2 and move 1 have outliers that outdo the highest in move 3.

The $\chi^2$ test (Table 13, p. 90) shows that the results indeed are significant ($p < 0.05$) but only for the new references, i.e. references published after 1995. The contingency table (Table 12, p. 89) show that the median tiered similarity for the new references is tier 7 for move 1, tier 7 for move 2 and tier 9 for move 3. The tiered similarity groups are no longer of equal size, as we are only looking at references in the introductions.

**Coded algorithmically.** Again, we find most references in move 1 and least in move 3 (Table 14, p. 91), the mean scores on the similarity scale seem to show that references in move 3 have the highest similarity, but as we can see on the boxplot (Fig. 12, p. 90) we have the same pattern as the paragraph-coded moves: the top 5 or 6 references are placed in moves 1 or 2, even though the mean value is better for move 3.

The higher similarity values for move 3 are significant ($p < 0.05$), as we see in the $\chi^2$ test (Table 16, p. 93), but still only for new references (i.e. published after 1995). The median tiers for new references on the tiered similarity scale are 7 for move 1, 8 for move 2 and 9.5 for move 3. Thus the difference between the moves is more accentuated here than using the paragraph-level coding; this difference is also visually discernable when comparing Fig. 11 (p. 89) with Fig. 13 (p. 92).

### 4.5 Discussion

As it turned out, the introduction is a good place to start looking for the most relevant references, not only do we find the highest co-citation counts here, but also the highest scores on the similarity scale, and the most references too.

**Moves.** Determining the move for every sentence did not produce significant results, but the other two methods proved that in move 3 a greater proportion of the references are similar to the citing article compared to the rest of the introduction, though especially move 1 holds some references that are very similar to the source article.
This could be due to a random disturbance in my sample, or it could be a general trend that authors start the introduction by referencing a few works that are very similar to the present paper. If the latter is the case, it would be interesting if it would be possible to ‘catch’ these references, perhaps by looking at repeated citations in the same article, as it is possible that these references are cited again in move 3, but such a study is outside the scope of this paper.

Judging from the results above it seems to be worthwhile to enrich current citation databases with metadata about the location of the reference in the citing text. Information on which section the reference is placed in could have great potential for citation analysis, and additional information on where in the section the reference is, would be a valuable enrichment.

As there does not seem to be much qualitative difference between assigning moves by paragraph-level or algorithmically, except the algorithmic technique seems to be slightly better, the rational solution would be use the algorithmic method, which can be automated as it does not need intellectual evaluation (cf. Small 1982, p. 289).

Suggestion. I would propose that citation databases become enhanced with the following metadata:

<table>
<thead>
<tr>
<th>Additional metadata for documents:</th>
<th>Additional metadata for references:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles and sequence of sections</td>
<td>Which section(s) it is located in</td>
</tr>
<tr>
<td>Number of paragraphs in each section</td>
<td>Paragraph-number(s) in that/those section(s)</td>
</tr>
</tbody>
</table>

That way we could calculate the relative position of references, both in relation to the entire document and in relation to the section it is located in. Additionally it would include information on how many times a reference occur in the citing document. This would allow for the implementation of various weighting schemes.

Perspectives. With the creation of citation indexes three new areas in information science were enabled (Nicolaisen 2003):

- Information seeking based on citation search strategies.
- Knowledge organization based on bibliographic coupling and co-citation analysis.
- Research evaluation based on citation counts.

If existing citation databases were enriched with the above mentioned metadata, it could increase the number of options in all three areas. The information seeker who has a source document that he finds relevant could specify that he wanted to retrieve documents that cited his source in the middle to end of the introduction, which would probably increase the precision of the search. Alternatively he could look for documents that cite two selected source documents in the same section or even paragraph, increasing the odds that they are juxtaposed in the citing document.

The possibilities for knowledge organization studies are evident; it would be possible to introduce weights enhancing the influence of co-citations where the two documents are cited in the same section or in the same or two adjoining paragraphs. It is also quite possible that it could increase the quality of co-citation studies to weigh co-citations where one or both citations come from the introduction of the citing document, but that remains conjectural at this stage.
The potential prospects for research evaluation are not quite as apparent, but qualitative studies on what kinds of references there are in the different locations in the citing document (cf. Peritz 1983), could combined with the proposed new metadata pave the way for broad quantitative studies on the kinds of influence a certain institute, journal or author have on the scientific community.

4.6 Conclusion

We will now return to the hypotheses stated in the introduction. This study has demonstrated that it is possible to infer something about a cited document’s similarity with the citing document from the location of the reference in the citing text, indeed there is a marked difference between documents cited in the three main sections of research articles (introduction, method and results & discussion). I have also demonstrated that documents cited in the last part of the introduction are significantly more like the citing article than documents cited elsewhere in the introduction.

The three move-specific hypotheses were also confirmed. As we progress through the introduction, the amount of references decreases from abundant to scarce, and conversely the cited documents become increasingly more similar to the citing article.

It must be noted that the confirmation of the hypotheses applies only as far as the subject of this study cover, i.e. contemporary research articles in international chemistry. If it can be established that the trends demonstrated in this paper are applicable on research articles in general, it would be most advantageous to include metadata on the textual location of references in citation databases.

References


Roald Erik Frøsig  Citation Classification Based on Genre


Appendices

Appendix A: Bibliographic Information for the 30 Source Articles


Appendix B: Move Structure Determined on Sentence Level

<table>
<thead>
<tr>
<th>Article#1</th>
</tr>
</thead>
</table>
| §1 | The ability to systematically manipulate the shapes of inorganic nanocrystals remains an important goal of modern materials chemistry. The shape and size of inorganic nanocrystals control their widely varying electrical and optical properties. One means of achieving shape control is by the use of a static template to enhance the growth rate of one crystallographic face over another. For example, two-dimensional films are obtained when there is favorable epitaxy on a substrate. Pyramidal “dots” are obtained if there is strain between the growing crystallite and the epitaxial substrate, as in the growth of InAs on GaAs§ and Ge on Si,§ or the growth of calcite on carboxyl-terminated lipid bilayers. §
| §2 | There are also many demonstrated cases of anisotropic inorganic nanocrystal growth in liquid media. The vapor–liquid–solid growth mechanism in which a solid rod grows out of a supersaturated droplet has been very successful in creating one-dimensional materials§ and has been applied to growth of (insoluble) nanorods in a liquid medium. Structural reaction media like regular or inverse micelles§ have also been used successfully. An electrochemical method yielded anisotropic gold particles. A pronounced effect of the surfactant/precursor ratio on particle shape was observed in the case of platinum nanocrystals. Recently, a new mechanism for the formation of complex nanostructures has been proposed, called “oriented attachment”. In this mechanism, nanocrystals can join together with complete crystallographic alignment and eliminate free surfaces. This mechanism can lead to the formation of many interesting new nanostructures. §
| §3 | In this paper, we demonstrate systematic variation of the shape of colloidal semiconductor nanocrystals, using thermal decomposition of organometallic precursors in a hot (300 °C) mixture of trioctylphosphine oxide and hexylphosphonic acid. As in the growth of spherical CdSe nanocrystals in hot trioctylphosphine oxide, the surfactants dynamically adsorb to the growing crystallites, allowing atoms to add and subtract for high crystallinity. This enables the growing crystallites to anneal, resulting in good crystallinity, while suppressing particle aggregation. It is important to note that the growth mode of the nanocrystals depends strongly upon monomer concentration. At low monomer concentration, Ostwald ripening occurs, and small nanocrystals dissolve at the expense of larger ones. Such slow growth conditions favor the formation of a spherical particle shape (least surface area). On the other hand, at high monomer concentration, relative differences between the growth rates of different faces can lead to anisotropic shapes. The relative growth rates of the different faces can be controlled by suitable variation of the ratio of trioctylphosphine oxide and hexylphosphonic acid. Using a combination of these parameters, we demonstrate the controlled formation of CdSe nanocrystals with rod, arrow, teardrop, and tetrapod shapes. This wide variation of unusual shapes provides important information about the growth of the nanocrystals. |

<table>
<thead>
<tr>
<th>Article#2</th>
</tr>
</thead>
</table>
| §1 | Labeling of biological molecules using fluorescent tags is a common and very useful practice in biological science. Fluorescent small molecules (organic dyes) are used in both single and multiplex detection approaches. However, tagging of biomolecules using organic fluorophores has significant limitations. Organic fluorophores tend to have narrow excitation spectra, and often exhibit broad emission bands with red tailing, which makes simultaneous quantitative evaluation of relative amounts of different probes present in the same sample difficult, due to spectral overlap. Furthermore, variation of the absorption and/or emission spectra of dye tagged bioconjugates requires the use of chemically distinct molecular labels with attendant synthesis and conjugation challenges. Nonetheless, the use of multiple dye labels has achieved a considerable level of sophistication, as demonstrated by recent flow cytometry work involving a three-laser system and eight-color-making scheme to simultaneously measure a total of 10 parameters on cellular antigens. §
| §2 | Colloidal semiconductor nanocrystals, often referred to as “quantum dots” or “QDs”, have the potential to overcome problems encountered by organic small molecules in certain fluorescent tagging applications by combining the advantages of high photobleaching threshold, good chemical stability, and readily tunable spectral properties. During routine preparation, their inorganic core is capped with an organic layer such as a trioctyl phosphate/trioctyl phosphate oxide mixture (TOP/TOPO). Substitution of this organic cap with other appropriate groups (cap exchange) permits tailoring of nanoparticles for efficient dispersal in various organic solvents. Replacing the TOP/TOPO cap with polar terminated groups allows dispersal of these nanoparticles in aqueous solutions with preservation of optical and electronic properties. |
Colloidal quantum dots used most often to date in fundamental or applied studies are spherical nanocrystals with core sizes that vary between 15 and 120 Å in diameter.\(^4\) Their optical and electronic properties are dominated by carrier confinement (electron/hole), which results in size dependence of their optical properties, including light absorption, PL, electroluminescence (EL), and cathodoluminescence (CL).\(^5\) Nanocrystals composed of ZnSe, CdS, CdSe, and CdTe have emission spectra that span the visible spectrum.\(^1,4\) Nanocrystals made of other materials (e.g., InAs, InP, and PbS) have photoluminescent emission that extends into the near-IR.\(^6\)

CdSe nanocrystals have broad absorption spectra with a first absorption peak (band edge) that progressively shifts to longer wavelengths with increasing particle size.\(^7\) Their PL spectra are narrow (full width at half-maximum, fwhm, of 25?40 nm), and their emission positions track the absorption band edge location, spanning the visible spectrum.\(^8\) These nanocrystals are thus tunable fluorophores with absorption characteristics that allow simultaneous excitation of different particle sizes at a single wavelength, while they exhibit a luminescent emission that spans a wide range of wavelengths in the visible spectrum.\(^9\) Overcoating the CdSe core with a wider band gap semiconducting material such as ZnS or CdS, a process based on the concepts of band-gap engineering used in electronics, permits passivation of surface states and reduces the leakage of excitons outside the core.\(^10\) This enhances the photochemical stability and resistance to photobleaching of these nanomaterials.\(^11\) Overcoating also improves the luminescence quantum yield substantially without affecting the location or spectral bandwidth of light emission, i.e., CdSe/ZnS nanoparticles have an emission fwhm of 30?45 nm.\(^12\)

The above QD luminescence properties suggest the possibility of considerable sensitivity in fluorescent detection applications.\(^13\) Figure 1 shows a plot of the dependence of integrated PL intensity versus dilution factor (or nanoparticle concentration), along with two representative emission spectra at the lowest examined concentrations of CdSe/ZnS QDs in hexane.\(^14\) Substantial PL integrated intensities with good signal-to-noise ratios and well-resolved emission spectra are measured for nanoparticle dispersions with concentrations smaller than 1 nmol of particles per liter.\(^15\) This is comparable to the sensitivity reached in similar conditions with the best conventional organic fluorophores.\(^16\)

Experiments involving labeling of metallic and/or magnetic nanocrystals having organic coating molecules on their surfaces have been reported,\(^17\) and there have been two important studies utilizing CdSe nanocrystals bound to biological molecules.\(^18\) In one study, an avidin/biotin binding scheme was employed to attach CdSe/CdS core/shell nanoparticles to actin fibers.\(^19\) These dots were capped with an additional thin layer of derivatized porous silica, to render them water compatible.\(^20\) The route to the final conjugate was complex and time-consuming, and yielded a product with a relatively low quantum yield.\(^21\) Furthermore, it must be carried out at extremely dilute conditions, which means that the amount of material obtained is inherently limited.\(^22\) The second study used CdSe/ZnS core/shell nanoparticles capped with mercaptoacetic acid groups.\(^23\) In this case, a conventional covalent cross-linking approach based on 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) condensation was used to conjugate the nanoparticles to immunoglobulin G (IgG).\(^24\) Successful preparation of QD/IgG conjugates was reported, albeit again with low PL quantum yields.\(^25\) A more recent study reported immobilization of thiol-terminated DNA segments (20?25-mers) on mercaptopropionic acid capped CdSe/ZnS nanocrystals.\(^26\) The resulting DNA-modified QDs were subsequently used to build hybrid assemblies with DNA-modified gold nanocrystals.\(^27\) In this case, binding of DNA segments to the QDs was driven by favorable interactions between thiol groups and the inorganic surface, similar to what has been reported for gold particles.\(^28\) In another report, interactions of polynucleotides, such as DNA, with quantum dots induced quenching in the luminescence emission of 40-Å CdS nanocrystals.\(^29\) Our attempts to covalently attach proteins to CdSe?ZnS nanoparticles capped with dihydroxylic acid groups using EDC were relatively unsuccessful.\(^30\) With either immunoglobulin G (IgG) or ovalbumin these experiments consistently resulted in the formation of a high level of aggregates.\(^31\) We therefore developed an alternative conjugation strategy, based on self-assembly utilizing electrostatic attractions between negatively charged ligand acid capped CdSe/ZnS quantum dots and engineered bifunctional recombinant proteins consisting of positively charged attachment domains (containing a leucine zipper) genetically fused with desired biologically relevant domains.\(^32\) The strategy of combining the use of alkyl-COOH capped CdSe/ZnS nanocrystals and two-domain recombinant proteins cloned with a highly charged leucine zipper tail offers several advantages.

1. The alkyl-COOH terminated capping groups, which permit dispersion of the nanocrystals in water solutions at basic pH, also provide a surface charge distribution that can promote direct self-assembly with other molecules that have a net positive charge.
2. The synthetic approach used to prepare the QDs can be easily applied to a number of different core/shell nanocrystals and extended to...
other combinations of semiconducting materials, II-VI and III-V, which can generate a group of fluorescent probes with tunable emission over a wide range of wavelengths.

2 This contrasts with the need for developing specific chemistry routes for each organic fluorescent dye case by case.

2 (3) The fusion protein approach provides a general and consistent way to prepare a wide selection of biological macromolecules amenable to alterations of the interaction domain, such as charge, size, stability to pH, and temperature.

3 This also allows control of the assembly of individual proteins, e.g., into monomers, dimers, and tetrarsers, a property that can be exploited in protein packing around the nanocrystals to form complex bioconjugates.

3 In the present report we describe the design, preparation, and initial analysis of a model QD/protein conjugate based on this strategy.

---

**Article#3**

§1

1 The palladium-catalyzed Suzuki cross-coupling of aryl halides and aryl triflates with arylboronic acids to form biaryl derivatives has emerged as an extremely powerful tool in organic synthesis.1

1 Compounds that contain a biaryl linkage have a diverse spectrum of applications, ranging from materials science4 to pharmaceuticals.

For example, with respect to pharmaceuticals, the biaryl group is a key feature in the sartan family of drugs for high blood pressure5 and in important natural products, such as the vancomycin antibiotics.5

1 Although the Suzuki cross-coupling is most commonly used to synthesize biaryls, reactions of vinyl halides/triflates with arylboronic acids, to generate styrene derivatives, are an important component of the broad utility of the process.

1 The innocuous nature of boronic acids, which are generally nontoxic and thermally, air-, and moisture-stable, is a practical advantage of the Suzuki reaction, relative to many other cross-coupling processes.

§2

1 There are a large number of parameters in a Suzuki reaction: palladium source, ligand, additive, solvent, temperature, etc. and there are, correspondingly, a large number of protocols for accomplishing the transformation, the choice of which depends on the structure of the reactants.1

2 Clearly, the development of a general protocol that could reliably effect cross-coupling, independent of the structure of the reactants, would be worthwhile.

§3

1 In 1998, we discovered that Pd2(dba)3/P(t-Bu)3 catalyzes the Suzuki cross-coupling of a wide array of aryl chlorides and arylboronic acids (eq 1).1

1 The high reactivity of this system toward typically unreactive aryl chlorides suggested to us that it might serve as a very versatile catalyst, under mild conditions,6 for Suzuki cross-couplings of other substrates.

3 In this report, we describe significant progress toward substantiating that possibility.

§4 (Results and Discussion)

Suzuki Cross-Coupling of Aryl Chlorides.

1 By way of background, at the time that we initiated our studies of palladium chemistry, one of the most important limitations of the Suzuki cross-coupling reaction was the poor reactivity of aryl chlorides, which are perhaps the most attractive family of aryl halide substrates due to their low cost and their ready availability.1

1 In fact, prior to 1998, reports of efficient palladium-catalyzed Suzuki couplings of aryl chlorides were limited to reactions of activated substrates (i.e., heteroaryl chlorides and aryl chlorides that bear an electron-withdrawing group), which generally only proceeded at high temperature (75°-130 °C).1,8

1 Since 1998, several research groups have described electron-rich ligands for palladium that overcome this limitation, specifically, arylidalkylphosphines (Buchwald;10,11 Bei and Guram12), P(t-Bu3)3 (Fu13), and N-heterocyclic carbenes (Nolan;14 Herrmann15).

2 With respect to Suzuki couplings of aryl chlorides that proceed at room temperature, the only successful catalyst systems reported to date are those of Buchwald (very general) and of Kocovsky (one example).16

§5

1 In our initial communication, we reported a general method for the Suzuki cross-coupling of aryl chlorides and arylboronic acids in the presence of a Pd2(dba)3/P(t-Bu)3 catalyst system, with Cs2CO3 as the base (80°-90 °C in dioxane; eq 1).1

1 Toward the end of that study, we discovered that CsF is also a highly effective base.17

2 We have since investigated the replacement of CsF by less expensive KF,12b,c and we have found that Suzuki reactions are even more rapid in the presence of KF.

2 The Pt(t-Bu)3:Pd is an important parameter: whereas use of a 1:1 ratio furnishes a very active catalyst, use of a 2:1 ratio leads to an extremely slow reaction.12

2 Among palladium sources, Pd2(dba)3 is superior to Pd(OAc)2.18

---

**Article#4:**

§1

1 Recently, we reported a new "programmed assembly" strategy19 for organizing nanoparticles into periodic functional materials.20

1 This strategy utilizes nanoparticle building blocks modified with alkanethiol-capped oligonucleotides (single-stranded DNA) and complementary linker oligonucleotide (DNA) strands to form extended structures with control over particle chemical composition, periodicity, and aggregate thermal stability, thereby providing control over the optical, mechanical, and electrical properties of these hybrid bioorganic materials.21

1 Importantly, we already have exploited the optical properties of one such system, formed from 13 to 17 nm diameter Au particles, in the development of a highly selective colorimetric diagnostic method for DNA.22

1 This diagnostic method relies on the distance-dependent optical properties of gold particles.23
Specifically, it is well documented that the plasmon frequency of a collection of gold particles can be controlled by adjusting the distance between discrete nanoparticles or layers of nanoparticles.

Based on these studies, we rationalized that it might be possible to control the optical properties of our composite materials by controlling the length of the DNA linker molecules/target molecules and the average distance between the particles, Scheme 1.

Moreover, this system poses an unprecedented opportunity to explore this relationship over a length scale that is broader and longer than the length scale associated with conventional organic interconnect molecules.

Stable oligonucleotide interconnects, in the 8 to 200 nucleotide range (27'680 Å), can be readily synthesized via standard solid-phase procedures.

Early solution and modeling studies on the optical properties of metal nanoparticle aggregates and agglomerates focused on relatively small structures (27'100 particles) with very small nanoparticle spacings (0).

The general conclusions from these studies are that for closely spaced metal nanoparticles, only small collections (27'100) of particles are required to give substantial shifts in the plasmon resonance.

Our DNA-based assembly method results in macroscopic assemblies consisting of thousands of particles and, therefore, offers a unique opportunity to study the collective properties of large nanoparticle network assemblies that have longer interparticle distances than previously studied systems.

Herein, we show that tailorability of the optical properties using this programmed assembly approach is indeed achievable through the use of different length DNA particle interconnects, but the observed optical effects in this system are as much a function of aggregate size (and the corresponding growth rate is dependent upon linker length) as interparticle distance.

This surprising result is significant since it impacts not only our understanding of the colorimetric response associated with our diagnostic method but also any effort aimed at using this approach to tailor the optical properties of these novel network materials.

The potential for use of two-photon absorbing molecules in applications ranging from optical limiting* to three-dimensional (3D) fluorescence microscopy and 3D microfabrication and optical data storage has stimulated research on the design, synthesis, and characterization of new molecules with large two-photon absorptivities.

The ability to achieve 3D spatial resolution arises from the fact that the rate of simultaneous absorption of two photons depends quadratically on the intensity of the incident laser light.

If a tightly focused beam is used, the intensity is highest at the focus and decreases quadratically with the distance (z) from the focal plane, for distances larger than the Rayleigh length.

Thus, the rate at which molecules are excited decreases very rapidly (as z-to-4) with the distance from the focus and the excitation is confined in a small volume around the focus (on the order of z%, where z is the wavelength of the incident beam).

The linear absorption of a material may be very weak in the wavelength range where the two-photon absorption occurs, providing the potential to excite materials at greater depth than might be possible via one-photon excitation.

Additionally, since the wavelength used for two-photon excitation is roughly twice that for one-photon excitation, the influence of scattering on the beam intensity is greatly reduced, by about a factor of 16 (since the light scattering efficiency scales as z-to-4).

These are clear advantages for applications such as imaging in absorbing or scattering media, like biological tissues.

Also, highly transparent nonlinear absorptive materials are of great interest in optical-limiting applications.

For many applications, chromophores with large two-photon absorptivities are advantageous, because a greater degree of excitation can be obtained or lower laser intensities can be used for pumping.

Thus, there is a need to develop detailed structure/property relationships for two-photon absorbing molecules in order to guide the design of molecules whereby one can systematically increase the two-photon cross section and tune the position of the two-photon absorption peak.

The development of such structure/property relationships will greatly benefit the design of new materials wherein improved two-photon absorption is combined with other molecular properties or processes (for example, large fluorescence quantum yield, efficient intersystem crossing, and low oxidation potential) that make those materials suitable for given applications.

We have previously reported on a design strategy for chromophores with large two-photon absorptivities.

One of these chromophore designs involves symmetrically substituted conjugated molecules with the general structure D-?-D, where D indicates an electron-donor group and ? is a conjugated bridge.

Experimental and theoretical evidence shows that the bis-donor substituted compounds can have ? values more than 1 order of magnitude larger than the corresponding unsubstituted molecules.

This enhancement in ? has been correlated with intramolecular charge transfer from the terminal donor groups to the ?-bridge.

The magnitude of ? can thus be controlled through the modification of the molecular structure in such a way as to affect the amount of intramolecular charge transfer.

Specifically, ? is increased when the conjugation length of the ?-system is increased (since the charge is transferred over a longer distance) and when electron acceptors (A) are attached as side groups to the ? bridge, to form molecules of general structure D-?-A-?-D (wherein the amount of charge transferred is increased).

Molecules based on the scheme A-?-D-?-A are also characterized by large ? (with peak absorptivities in the red and near-infrared spectral range).

This result provides evidence that the charge transfer is effective in enhancing the two-photon absorptivity irrespective of the direction of the transfer (from the ends to the center of the molecule or from the center to the ends).
It is known that for linear polyenes and several classes of \(1,2\)-substituted polyenes with more than two or three double bonds, the lowest excited singlet state is the 2Ag state, while the one-photon-allowed 1Bu state lies at slightly higher energy.\(^*\)

This finding allows one to explain the absorptive and emissive characteristics of these molecules.

We will show that the bis-donor substituted diphenylpolyenes we have prepared exhibit a different ordering of the excited states for medium-length molecules; specifically the 1Bu state is energetically below the 2Ag state for \(m < 5\) (\(m\) = number of double bonds in the polyene).

We present here the results of studies aimed at understanding the two-photon spectroscopy, the ordering of electronic states, and the photophysics of D-?-D systems in which the conjugated bridge is a diphenylpolyene or a phenylene-vinylene-type chain.

The issues addressed in this paper are as follows:

(i) The two-photon-induced fluorescence excitation spectra for a series of systematically varied structures will be presented.

(ii) The effect of pulse duration on the cross sections, as measured by two-photon-induced fluorescence excitation spectroscopy, using both nanosecond and picosecond laser pulses, will be examined for a series of substituted diphenylpolyenes.

(iii) The effect of chain length, dialkylamino or diarylamino terminal groups, and type of conjugated bridge on the position and magnitude of the lowest energy two-photon absorption band will be discussed.

(iv) The ordering of the lowest energy one- and two-photon-allowed excited states for these D-?-D chromophores will be discussed in comparison with similar conjugated molecules.

(v) A study of the fluorescence lifetimes following either one-photon or two-photon excitation will be presented and their relationship to the ordering of the 1Bu and 2Ag states will be discussed.

(vi) An examination of how the transition dipole moments between the ground state and the one-photon excited singlet state and between the one- and two-photon excited states vary with chain length, with type of bridge, and the presence of terminal donor groups will be presented.

Transition moments determined from experimental spectroscopic parameters using a three-level model will be compared with those from quantum chemical calculations.

**Article#6:**

§1 Werner complexes, \(\text{M}^2\text{L}_4\text{NCS}\) \((\text{M} = \text{Ni}^{2+}\) or \(\text{Co}^{2+}\)\), Prussian blue compounds, \(\text{Fe}_{4}\text{[Fe(CN)6]}\) \((x = 14\text{?}16)\),\(^*\) and Hofmann clathrates and their derivatives, \(\text{Cd(NH}_3\text{)}_2\text{Ni(CN)}_4\) \((G = \text{guests such as C}_6\text{H}_6\) and C\(_{12}\text{H}_{10}\))\(^\text{**}\), are widely known as materials that can reversibly sorb small molecules.

Open frameworks are also known to exist in numerous other coordination compounds such as copper(I) bis(adiponitrilo) nitrate and Cu\(_3\text{en}_2\text{CN}_4\) \((\text{en} = \text{ethylenediamine})\), however, their sorption behavior was not reported.\(^*\)

Nevertheless, these compounds have inspired the current interest in the copolymerization of a wide range of larger organic linkers with transition metals to yield new classes of designed metal/organic frameworks (MOF) having unusual open architecture and unprecedented pore size and shape.\(^*\)

However, these frameworks are commonly plagued by their inability to support porosity: the removal of guests, incorporated during synthesis, or their exchange leads to collapse and decomposition of the framework, thus precluding study of their inclusion chemistry - an aspect that is critically important for their utility as porous materials.

§2 Secondary Building Units (SBUs).

Our approach to the preparation of robust frameworks has been to employ bis- and tris-bidentate carboxylate linkers such as 1,4-benzenedicarboxylate (BDC) and 1,3,5-benzenetricarboxylate (BTC), which offer important advantages due to their rigidity and consequent tendency to form rigid metal carboxylate clusters that ultimately act as secondary building units (SBUs) in the extended solid.

In fact, the extended structure analogues of molecular dinuclear, trinuclear and tetranuclear zinc(I) carboxylate clusters (Figure 1)\(^*\) have been successfully prepared using BDC and BTC as linkers to give Zn(BDC)\(_2\) (DMF)\(_2\) (H\(_2\)O) (DMF = N,N'-dimethylformamide) (MOF-2), Zn\(_3\) (BDC)\(_3\) (6CH\(_3\)OH) (MOF-3), Zn\(_2\) (BTC)\(_2\) (NO\(_3\)) (C\(_2\)H\(_5\)OH) (MOF-4), and Zn\(_4\) (BTC)\(_3\) (DMF)\(_2\) (C\(_6\)H\(_5\)CN) (MOF-5).

The syntheses and atomic structures of these crystals have been reported and discussed in preliminary accounts.\(^*\)

The main focus of this report is to demonstrate, using sorption measurements, that such metal/organic frameworks can support permanent porosity and exhibit inclusion properties uncommon to traditional inorganic porous materials.

**Article#7:**

§1 Surface plasmon resonance (SPR) is a surface-sensitive analytical technique based on the ability to detect dielectric constant changes induced by molecular adsorption at a noble metal film.\(^*\)

In a typical experiment the reflectivity of a 50 nm thick Au film is measured as a function of incident angle.

The first commercial instrument (BIACore) to detect these changes by monitoring the formation of a biomolecular complex was introduced in 1990.\(^*\)

Since then, several systems have been developed, and this technique has been widely used in biomolecular interaction studies, including determination of affinity constants and kinetic binding parameters.\(^*\)

Unfortunately, the inability of conventional SPR to measure extremely small changes in refractive index hinders its application in ultrasensitive detection.

To address this drawback, several approaches have been developed.

Among them, substantial interest has been focused on utilizing external labels to enhance the sensitivity of the current technique.\(^*\)

In these approaches, a “sandwich” geometry is used, in which a surface-bound molecule (e.g. antibody or oligonucleotide) binds the analyte specifically, bringing it to the surface.
The event leads to a very small change in SPR response.
1. To increase the magnitude of the response, a second specifically interacting biomolecule is introduced.
2. Often this second, "probe", molecule carries a high molecular weight or high refractive index tag.
3. Consequently, a much larger change in SPR reflectivity can be observed.
4. To date, liposomes, \(^{15}\) latex particles, \(^{18,19}\) and certain proteins \(^{20}\) have been tested as amplification tags.

§3
1. Metallic nanoparticles have also been employed to enhance SPR response.\(^{21,22}\)
2. These particles offer ease of preparation, high density, large dielectric constant, and biocompatibility.\(^{23}\)
3. Pollard-Knight and co-workers used 30-nm diameter colloidal Au nanoparticles in an immunoassay to demonstrate the concept of this amplification technique.\(^{24}\)
4. Particle-enhanced SPR has also been used to develop cellular assays for cholera cells.\(^{25}\)
5. Natan and co-workers have developed an Au-amplified SPR sandwich immunoassay and achieved a picomolar detection of human immunoglobulin G.\(^{26}\)
6. The Natan group has also conducted systematic studies of the effect of particle size and surface coverage on SPR response.\(^{27,28,29}\)

§4
1. Herein, we further extend the scope of Au nanoparticle-amplified SPR to analysis of DNA hybridization.
2. Detection of DNA hybridization holds great promise for the rapid clinical diagnosis of genetic diseases.
3. SPR has been used in DNA analysis for nearly a decade.\(^{30,31}\)
4. However, few efforts have been devoted to improving sensitivity, despite intensive interest in kinetic studies of DNA-related interactions.
5. In this report, we describe the use of biocomplexes comprised of oligonucleotide probes and colloidal Au particles for DNA hybridization detection via SPR.
6. Specifically, oligonucleotide probes were conjugated to colloidal Au particles and used to selectively recognize surface-confined target DNA via sequence-specific hybridization with in situ detection.
7. A substantial improvement in the SPR response was achieved compared to the unamplified detection event.
8. To further support the validity of this approach, the melting behavior of the surface-bound DNA duplex was examined.
9. In addition, a restriction enzyme that can recognize a specific DNA duplex sequence was employed to release the particles from the surface.
10. Using a home-built SPR imaging instrument, a quantitation limit of less than \(8 \times 10^8\) molecules/cm\(^2\) for a 24-mer oligonucleotide and approximately 5 orders of magnitude in dynamic range were observed.
11. This result represents a greater than 1000-fold improvement over the unamplified binding event under similar conditions.\(^{32}\)

§5
1. To our knowledge, these experiments comprise the first example of Au-amplified SPR for DNA hybridization detection.
2. The sensitivity of this method already approaches that of standard fluorescence techniques for scanning gene chips.\(^{33}\)
3. Further improvements in the sensitivity of particle-amplified SPR can be achieved by optimizing hybridization conditions and reducing the background caused by nonspecific interactions.
4. In addition, larger particles can be used to greatly increase sensitivity, as previously demonstrated.\(^{34}\)
5. Thus, particle-amplified SPR offers a potentially powerful new approach to high throughput analysis of DNA arrays, such as those used in DNA sequencing and gene expression investigations, and may be applicable to polymerase chain reaction (PCR)-free DNA detection.
The molecular frameworks Cd(tcm)[Br(OMe)4]·cMeOH, Ni(otp)[NO3]2, [(Ni(bpy)(ArCOO)2)(MeOH)2]n, [(Zn(sala)(H2O)2)2]2H2O, Zn2(btc)(NO3)2·H2O·5EtOH, and Agthat)CO3·3CH3NO2 are all chiral, the last two displaying the helical (10,3)-a network, and retaining crystallinity on solvent removal.

A high degree of robustness to solvent removal has been reported recently in coordination frameworks incorporating chirally functionalized ligands.

Section 3

In this paper the combination of alcohol and pyridine equatorial ligands at an octahedral metal center connecting two trans diaxial btc ligands is shown to be a flexible recipe to prepare a wide range of chiral porous solids with large extracrystalline void regions.

The neutral M3(btc)2 framework adopts the (10,3)-a network (also known as Y*, SrSi2, or “Laves Graph”, one of the seven networks of equivalent three-connected points that repeat regularly in space and have the property that the shortest circuit from any point and including any two of the links meeting at the point is a decagon).

The extent of network interpenetration is controlled by whether the alcohol ligands bound equatorially to the metal are uni- or bidentate.

This network is intrinsically chiral due to the helices from which it is constructed, making the large void volume of this structure of particular interest.

Section 4

Bulk homochiral samples of enantiopure networks can be prepared from resolved template molecules with the hand of the helix directly controlled by that of the diol bound to the metal.

Careful desolvation below 145 °C affords crystalline samples with both empty chiral pores and vacuum coordination sites around the metal centers.

Move Article #9:

§1

Since the discovery of surfactant-templated silica mesostructures, the development of organic modification schemes to impart functionality to the pore surfaces has received much attention.

Most recently, using the general class of compounds referred to as bridged silsesquioxanes (RO)3Si?R'?Si(OR)3 (Scheme 1), three research groups have reported the formation of a new class of poly(brided silsesquioxane) mesophases (BSQMs) with integral organic functionality.

In contrast to previous hybrid mesostructures in which organic ligands or molecules are situated on pore surfaces, this class of materials necessarily incorporates the organic constituents into the framework as molecularly dispersed bridging ligands.

Although it is anticipated that this new mesostructural organization should result in synergistic properties derived from the molecular-scale mixing of the inorganic and organic components, few properties of BSQMs have been measured.

In addition, samples prepared to date have been in the form of granular precipitates, precluding their use in applications such as membranes, fluidics, and low-k dielectric films needed for all foreseeable future generations of microelectronics.

Capacitance-voltage measurements along with a variety of structural characterization procedures were performed to begin to elucidate structure-property relationships of this new class of thin-film and particulate mesophases.

§2

M The initially reported syntheses of BSQMs involved precipitation in basic aqueous media followed by aging for periods up to several days.

M To form homogeneous films, precipitation must be avoided, and the self-assembly process must be accomplished within the brief time span of the dip- or spin-coating operation (at most several seconds).

M To meet these requirements, we prepare dilute, homogeneous, ethanolic solutions that suppress silica/surfactant self-assembly and employ acidic conditions that retard siloxane condensation.

M In a typical synthesis procedure, requisite amounts of 1 (from Gelest and triply distilled prior to use), 2 (synthesized according to ref 15), or 3 (synthesized according to ref 15) are dissolved in ethanol followed by addition of 1/38 wt % surfactant [cationic (CTAB, CH3(CH2)15N+(CH3)3Br-), nonionic (Brij-56, C16H33(OCH2CH2)10OH), anionic (SDS, C12H25SO3Na+), or block copolymer (P123, H2O(CH2CH2)x(0<x<20)(OCH(CH2CH2)2OH)]) and an aqueous solution of HCl (0.1?1.0 N).

M The investigated range of starting compositions is represented by the molar ratios Si:EtOH:H2O:HCl:surfactant = 1:2-5:0.004:(0.05-7.0)-18.

M To evaluate the effect of substitution of bridged silsesquioxanes for siloxanes on the properties of resultant thin-film mesostructures, a series of films was prepared with varying ratios of TEOS and (EtO)3Si(CH2)27Si(0Et)3.

M For this series, the starting compositions are represented by the molar ratios TEOS:1 (n = 2):EtOH:H2O:HCl:surfactant = (0.3?3):1.0-25.0:0.44:0.8-0.0084.

M In all cases the initial surfactant concentration (c0) is much less than the critical micelle concentration (cmc), implying mesoscale homogeneity of the starting sols.

§3

M Films were prepared by spin- or dip-coating, and nanoparticles were prepared by an aerosol-assisted self-assembly process described previously.

M In both cases, preferential ethanol evaporation concentrates the sol in water, nonvolatile surfactant, and organically bridge polysilsesquioxane species.

M By choosing the initial acid concentration to retard the competing process of siloxane condensation, we exploit the progressively increasing surfactant concentration to drive self-assembly of polysilsesquioxane/surfactant micelles and their further organization into liquid crystalline mesophases.

M Pre-existing, incipient polysilsesquioxane/surfactant mesostructures (which exist at solid?liquid and liquid?vapor interfaces at c < cmc) serve to nucleate and orient the mesophase development.

RD The result is rapid (several seconds) formation of thin-film or particulate BSQMs that are oriented with respect to the solid?liquid and/or...
We observe a consistent trend of increased modulus and hardness and decreasing dielectric constant with substitution of the bridged phenylene bridge and 29Si resonances at δ71.0 and δ78.2 ppm, corresponding to T2 and T3 species, respectively (where the superscript denotes the number of bridging oxygens covalently bonded to the trifunctional silicon center).

Based on a cylindrical pore model, we used the ratio of the thickness-normalized pore volume and surface area to calculate hydraulic pore diameters of 1.8 and 2.5 nm, respectively, for the CTAB and Brij-56 films.

Retention of the bridging organic ligands in the calcined specimens was confirmed using 13C and 29Si MAS NMR spectroscopy. For powders prepared from 1 with 6 wt % CTAB or 4 wt % Brij-56, or SDS surfactant before and after calcination at 250 °C in N2.

Interestingly, we observe undulations and defects in the stripe pattern in a plane normal to the substrate surface, whereas for hexagonal thin-film mesophases prepared from TEOS these features are observed only in planes parallel to the substrate surface.

Figure 2B shows a cross section of the film prepared with 4 wt % Brij-56.

Based on the hexagonal pattern of spots and the 5.3-nm spot-to-spot spacing, we interpret the upper right region as a [111] orientation of a face-centered cubic structure with a δ9.1-9.3 nm, consistent with XRD results and previous cubic thin-film mesophases prepared from Brij-56 and TEOS.

Figure 2C shows the TEM micrograph of a calcined spherical nanoparticle prepared using the phenylene-bridged precursor 2 with 8 wt % P123 block copolymer surfactant.

Figure 2D shows the TEM micrograph of a calcined spherical nanoparticle prepared using precursor 3 with 8 wt % P123 block copolymer surfactant.

This multilamellar vesicular structure is composed of concentric bridged polysilsesquioxane layers containing integral vinyl functionality.

The porosity of the thin-film specimens from 1 with 6 wt % CTAB or 4 wt % Brij-56 was characterized by nitrogen sorption measurements performed on film specimens using a surface acoustic wave (SAW) technique.

The SAW N2 sorption isotherms shown in Figure 3 are qualitatively consistent with those previously obtained for the corresponding TEOS-derived films and typical of surfactant-templated mesophases in general. Based on a cylindrical pore model, we used the ratio of the thickness-normalized pore volume and surface area to calculate hydraulic pore diameters of 1.8 and 2.5 nm, respectively, for the CTAB and Brij-56 films.

Retention of the bridging organic ligands in the calcined specimens was confirmed using 13C and 29Si MAS NMR spectroscopy. For powders prepared from 1 with 6 wt % CTAB, we observed a large 13C resonance at 5.2 ppm, corresponding to the ethylene bridge, and a very broad 29Si resonance at δ78.5 ppm, corresponding to trifunctional (T) silicons.

There was no evidence of tetrafunctional Q silicons at δ701 or δ708 ppm, implying complete retention of the bridging ligands. For the phenylene-bridged system prepared from 2 with 8 wt % P123, we observed a 13C resonance at 133.4 ppm corresponding to the phenylene bridge and 29Si resonances at δ71.0 and δ78.2 ppm, corresponding to T2 and T3 species, respectively (where the superscript denotes the number of bridging oxygens covalently bonded to the trifunctional silicon center).

Again, there was no evidence of Q silicons. To begin to establish structure?porosity relationships, we prepared a series of films from TEOS and 1 (n = 2) with molar ratios TEOS:1 = 75:25 (TB1), 50:50 (TB2), and 25:75 (TB3).

Synthesis and processing procedures were chosen to create isotropic disordered (wormlike) thin-film mesophases17 with comparable film thicknesses (measured by spectroscopic ellipsometry) and porosities (measured by analyses of SAW-based N2 sorption isotherms). After calcination at 350 °C under N2 to remove the surfactant templates, all films were vapor-treated with hexamethyldisilazane to avoid water adsorption.

Table 1 compares values of the dielectric constants (measured using a standard capacitance/voltage technique employing a mercury probe), Young's modulus, and hardness (modulus and hardness calculated from nanoindentation measurements at a constant indentation depth, assuming a Poisson ratio of 0.2).

We observe a consistent trend of increasing modulus and hardness and decreasing dielectric constant with substitution of the bridged

| RD | Through variation of the shape of the surfactant along with its charge and initial concentration, this evaporation-induced self-assembly route can be used to attain a range of thin-film or particulate mesophases. |
| RD | The major XRD peak in the Brij-56 system may be interpreted as a (200) reflection of a cubic mesophase with a = 9.2 nm. |
| RD | TEM micrographs of calcined thin-film specimens evaluated by XRD are shown in Figure 2. |
| RD | Regions adjacent to the particle circumference, which correspond to the thinnest regions of the specimen in the imaging direction, exhibit a hexagonal close-packed texture. |
| RD | We interpret this structure as a hexagonal mesophase composed of close-packed cylindrical pore channels aligned parallel to the particle circumference. |
| RD | This orientation arises from the initial nucleation of the hexagonal mesophase at the air/liquid interface and its growth inward, driven by an evaporation-induced radial concentration gradient. |
| RD | We note that particles prepared from the comparable TEOS/8 wt % P123 system exhibit a multilamellar vesicular mesostructure. |
| RD | The porosity of the thin-film specimens from 1 with 6 wt % CTAB or 4 wt % Brij-56 was characterized by nitrogen sorption measurements performed on film specimens using a surface acoustic wave (SAW) technique. |
| RD | The SAW N2 sorption isotherms shown in Figure 3 are qualitatively consistent with those previously obtained for the corresponding TEOS-derived films and typical of surfactant-templated mesophases in general. |
| RD | Based on a cylindrical pore model, we used the ratio of the thickness-normalized pore volume and surface area to calculate hydraulic pore diameters of 1.8 and 2.5 nm, respectively, for the CTAB and Brij-56 films. |
| RD | Retention of the bridging organic ligands in the calcined specimens was confirmed using 13C and 29Si MAS NMR spectroscopy. For powders prepared from 1 with 6 wt % CTAB, we observed a large 13C resonance at 5.2 ppm, corresponding to the ethylene bridge, and a very broad 29Si resonance at δ78.5 ppm, corresponding to trifunctional (T) silicons. |
| RD | There was no evidence of tetrafunctional Q silicons at δ701 or δ708 ppm, implying complete retention of the bridging ligands. |
| RD | For the phenylene-bridged system prepared from 2 with 8 wt % P123, we observed a 13C resonance at 133.4 ppm corresponding to the phenylene bridge and 29Si resonances at δ71.0 and δ78.2 ppm, corresponding to T2 and T3 species, respectively (where the superscript denotes the number of bridging oxygens covalently bonded to the trifunctional silicon center). |
| RD | Again, there was no evidence of Q silicons. |
| RD | To begin to establish structure/porosity relationships, we prepared a series of films from TEOS and 1 (n = 2) with molar ratios TEOS:1 = 75:25 (TB1), 50:50 (TB2), and 25:75 (TB3). |
| RD | Synthesis and processing procedures were chosen to create isotropic disordered (wormlike) thin-film mesophases17 with comparable film thicknesses (measured by spectroscopic ellipsometry) and porosities (measured by analyses of SAW-based N2 sorption isotherms). |
| RD | After calcination at 350 °C under N2 to remove the surfactant templates, all films were vapor-treated with hexamethyldisilazane to avoid water adsorption. |
| RD | Table 1 compares values of the dielectric constants (measured using a standard capacitance/voltage technique employing a mercury probe), Young's modulus, and hardness (modulus and hardness calculated from nanoindentation measurements at a constant indentation depth, assuming a Poisson ratio of 0.2). |
| RD | We observe a consistent trend of increasing modulus and hardness and decreasing dielectric constant with substitution of the bridged |
Roald Erik Froisig

Citation Classification based on Genre

Appendix B

Move

Article#10:

§1

1 Residual dipolar couplings measured in dilute liquid crystalline phase provide important restraints for molecular structure determinations by NMR spectroscopy which cannot be obtained otherwise.

1 Several different liquid crystal media have been proposed to achieve the necessary partial alignment of the molecules with respect to the magnetic field, including phospholipid bicelles composed of mixtures of DMPC/DHPC or DODPC/CHAPSO, and lamellar or nematic phases made of cetylpyridinium chloride/hexanol mixtures, filamentous phages or purple membrane fragments in aqueous solution.

2 None of these media, however, is universally applicable.

2 DMPC/DHPC mixtures suffer from irreversible degradation at low pH due to hydrolysis, DODPC/CHAPSO mixtures are more stable at low pH but do not form a stable liquid crystalline phase at neutral pH, cetylpyridinium chloride/hexanol requires high salt concentrations for formation of a stable phase which is destabilized by high protein concentrations, cetylpyridinium bromide/hexanol seems to work best under low ionic strength conditions, phages work well at neutral pH but tend to aggregate below pH 6.5 and purple membrane fragments bind proteins rather tightly, resulting in line broadening and orientation tensors which do not merely reflect the shape of the solute molecule.

3 Here we propose the application of an alternative set of liquid crystalline media composed of n-alkyl-poly(ethylene glycol)/n-alkyl alcohol mixtures.

3 These media are uncharged, insensitive to pH, fairly insensitive to salt, feature little or no binding capacity for macromolecules, and can be used at temperatures below 0 °C up to almost 40 °C, covering the temperature range of interest for most NMR studies of biological macromolecules.

3 In addition, the technical product "glucopone", which contains different n-alkylated carbohydrates, is shown to form a lamellar phase in mixture with n-hexanol which is suitable for the measurement of residual dipolar couplings.

§2

1 Different alkyl-poly(ethylene glycol) molecules are denoted as CmEn, where m is the number of carbons in the n-alkyl group and n is the number of glycol units in the poly(ethylene glycol) moiety.

1 Phase diagrams have been reported for ternary mixtures of C12E5, n-hexanol, and H2O, and mixtures of C8E5, n-octanol, and H2O.

2 In the present work, we also explored mixtures of C12E6, n-hexanol, and H2O.

2 Under certain conditions, all these systems form a lyotropic liquid crystalline phase referred to as L?, where the hydrophobic n-alkyl chains aggregate into planar bilayers with the hydrophilic poly(ethylene glycol) headgroups directed into the water phase.

2 The L? phase is optically clear with slight opalescence, forming a lamellar-like superstructure where the spacing of the stacked bilayers and thus the degree of alignment of guest molecules in the water-rich layers can be tuned by the surfactant/water ratio.

2 The spacing can range from several hundred nanometers (1 wt %) to a few nanometers in the more concentrated regime (>5 wt %).

2 In a magnet, the bilayer surfaces are oriented parallel with respect to the direction of the magnetic field, presumably with a superstructure in which the bilayers bend into a set of concentric tubes of different radii with the axis aligned along the magnetic field.

2 At high surfactant concentration (>50%) and high temperatures (>50 °C), liquid crystalline phases are already formed by binary mixtures without alcohol.

2 The addition of n-alkyl alcohols lowers the temperature range of stability for the liquid crystalline phase and, most important, makes the L? phase accessible even at low surfactant concentrations.

§3

3 The bilayers formed by C12E5/n-hexanol, C8E5/n-octanol, and C12E6/n-hexanol carry an effective poly(ethylene glycol) (PEG) coating that provides a means of minimizing protein adsorption, in the same way as PEG-derivatized surfaces are suitable for protein chromatography and enhanced biocompatibility of implants, and PEG is used as a cosolvent in protein crystallization.

3 Interestingly, PEG groups can also physically stabilize bilayers, as the addition of 1% PEG2000–PE increases the stability of DMPC/DHPC bicelles at surfactant/H2O ratios above 10 wt %.

3 In the present work, we have explored the stability range of the L? phase formed by different CmEn/alcohol systems in the concentration range of 3–8 wt %, where residual dipolar couplings are easily measured for one-bond couplings in proteins, without resulting in excessive line broadening due to 1H–1H couplings.

Move

Article#11:

§1

1 DNA-modified surfaces are the subject of considerable current activity in the field of biotechnology.

1 Despite their growing importance, several aspects of the performance of these novel composite materials is far from optimum, and their surface chemistry remains poorly characterized.

2 Desired attributes of DNA-modified surfaces include the following: (a) surface flatness and chemical homogeneity; (b) ability to control surface chemical properties such as polarity or hydrophobicity, which impact strongly upon nonspecific binding properties; (c) amenability to DNA hybridization (duplex formation) and enzymatic manipulation with DNA-modifying enzymes such as ligase, polymerase, and restriction enzymes; (d) ability to control DNA surface density; (e) thermal and chemical stability; and (f) reproducibility of preparation.

2 Few if any of these criteria are met by the surface chemistries presently in use.

§2

2 Remarkable advances have been made in microelectronics technology over the last 20 years, primarily due to increasingly powerful
A similar trend has become evident in the fields of biology/biotechnology, where arrays of tens of thousands of distinct DNA molecules on planar substrates have proven useful for the parallel analysis of genetic variation and gene expression levels. The development of robust, well-characterized surfaces and surface attachment strategies for biological analyses could benefit greatly from the well-developed infrastructure that exists in microelectronics.

Previous researchers have successfully attached DNA to substrates such as latex beads, polystyrene, optical fibers, carbon electrodes, gold, and oxidized silicon. Although largely unexplored to date, crystalline silicon is a particularly attractive alternative substrate for DNA immobilization because of its purity and well-defined structure, and because crystalline silicon substrates provide the opportunity to take full advantage of existing technologies in microelectronics processing.

Chemical functionalization of silicon is complicated by the fact that silicon readily oxidizes in air to produce an oxide that is chemically similar to glass. Unfortunately, the use of glass or oxidized silicon substrates for highly parallel DNA attachment presents some problems that are a direct consequence of the fact that glass and oxidized silicon surfaces are amorphous, and that the relative number of Si–O–Si and Si–OH linkages exposed at the surface is highly dependent on the past history of the sample. This irreproducibility in surface chemistry leads to difficulties in the control, reproducibility, homogeneity, and stability of subsequent DNA attachment.

Planar glass and oxidized silicon substrates can be modified using methoxysilane or chlorosilane reagents to attach organic functional groups such as amines or thiols to serve as DNA attachment sites. Silane reagents containing only a single surface reactive functionality (e.g., monochloro- or monomethoxysilanes) produce poorly stable films. The stability can be improved by using multiple reactive groups (e.g., trichloro- or trimethoxysilanes) to form more Si–O–Si linkages, but such polyfunctional reagents also lead to poorly controlled surface polymerization reactions, compromising surface homogeneity and reproducibility.

A solution to this problem would be to develop strategies for direct attachment of DNA to silicon surfaces without an intervening oxide layer. In recent years, new attachment methods for the organic functionalization of silicon surfaces through formation of direct silicon–carbon bonds have been reported. In this report we utilize these reactions as a route for the attachment and hybridization of DNA to crystalline silicon substrates. The resultant DNA-modified silicon surfaces are reproducibly prepared, stable to the conditions of DNA hybridization, and show no detectable nonspecific binding.

The high fidelity of these surfaces promises substantial benefit to the emerging technology of large-scale biological analysis using nucleic acid arrays.
Re6 clusters has been generated. In addition, a range of new layered and porous structures containing Mn2+, Fe2+, Co2+, Ni2+, Zn2+, and Cd2+ ions spaced apart.

\[ \text{[Mn}_2\text{(H}_2\text{O})_4\text{][Ru(CN)}_6\text{]·4H}_2\text{O}} \]

Thus far, the metal-cyanide frameworks of Prussian blue (Fe4[Fe(CN)_6]3·14H_2O), the greater size of the cluster ions, however, suggests that their substitution into known cyanometalate frameworks of an Re6 octahedron with each face capped by a Q atom and a terminal cyanide ligand extending from each Re apex.

\[ \text{[(Ph}_3\text{P})_2\text{N}^+\text{][Ru(bpy)(CN)}_4\text{]2} \]

Indeed, this type of selectivity is a distinct advantage of extended framework sensing materials. Encapsulation of the dye Nile Red in zeolite Y for example, the [(Ph3P)2N]+ salt of the solvatochromic [Ru(bpy)(CN)4]2− complex can function as a humidity sensor.

Other approaches have focused on incorporating established solvatochromic molecules or ions into solid materials.

For example, the [(Ph3P)2N]+ salt of the solvatochromic [Ru(bpy)(CN)4]2− complex can function as a humidity sensor.

Encapsulation of the dye Nile Red in zeolite Y or incorporation of luminescent Tb3+ ions into the framework of a microporous solid permit shape and size selective sensing.

Indeed, this type of selectivity is a distinct advantage of extended framework sensing materials.

Resembling enlarged cyanometalate ions, the recently prepared [Re6O8(CN)_6]^3− (Q = S, Se, Te) clusters exhibit a geometry consisting of an Re octahedron with each face capped by a \(\beta\)-Q atom and a terminal cyanide ligand extending from each Re apex.

The greater size of the cluster ions, however, suggests that their substitution into known cyanometalate-based solids could enlarge extant framework cavities, leading to materials with enhanced inclusion properties.

Thus far, the metal-cyanide frameworks of Prussian blue (Fe4[Fe(CN)_6]3·14H2O), Na2Zn3[Fe(CN)_6]2·9H2O, and [Mn2(H2O)4][Ru(CN)_4]·4H2O have all been successfully expanded using just this strategy.

In addition, a range of new layered and porous structures containing Mn2+, Fe2+, Co2+, Ni2+, Zn2+, and Cd2+ ions spaced apart by the Re6 clusters has been generated.
Frequently, the metal ions incorporated in these open framework structures exhibit one or more coordinated water molecules that can be removed by thermolysis to produce accessible, vacant metal coordination sites.

Herein, we demonstrate how Co2+ ions, well-known for the dramatic color changes associated with their facile interconversion between octahedral and tetrahedral coordination geometries, can be incorporated into porous solids to produce materials capable of detecting and identifying volatile organic compounds.

In practice, both techniques have specific drawbacks. These features are usually known for the dramatic color changes associated with their facile interconversion between octahedral and tetrahedral coordination geometries, can be incorporated into porous solids to produce materials capable of detecting and identifying volatile organic compounds.

In such a device configured, the compound is not only able to generate electrons and holes under light irradiation but it also provides pathways for their subsequent collection at opposite electrodes and a photocurrent is obtained. To avoid any problems arising from bad contacts at the junction, we have recently shown that the bicontinuous network can be chemically linking a hole-conducting oligophenylenevinylene (OPV) moiety to an electron-conducting fullerene subunit.

The C60/3PV hybrid compound has been incorporated in a photovoltaic cell constructed by spin-casting the compound on a glass substrate covered with indium-tin oxide (ITO) and depositing an aluminum film on top. In such a device configuration, the compound is not only able to generate electrons and holes under light irradiation but it also provides pathways for their subsequent collection at opposite electrodes and a photocurrent is obtained.

We now report in more detail the preparative procedures used to obtain C60/3PV as well as the synthesis of the corresponding higher homologues C60/4PV.

We also discuss their electrochemical and excited-state properties in solution using the related OPV derivatives 3PV and 4PV and the fulleropyrrolidine FP as reference compounds.

The observation of reversible, metastable photoinduced electron transfer from conjugated oligomers and/or polymers to C60 has attracted fair attention in view of its potential application in photovoltaic devices.

The performance of this type of device is very sensitive to the morphology of the blend. Ideally (to ensure efficient exciton dissociation), an acceptor species should be within the exciton diffusion range from any donor species and vice versa.

Moreover, both the donor and the acceptor phases should form a bicontinuous microphase separated network to allow bipolar charge transport. However, the donor and acceptor molecules are usually incompatible and tend to undergo uncontrolled macrophase separation.

To avoid any problems arising from bad contacts at the junction, we have recently shown that the bicontinuous network can be obtained by chemically linking a hole-conducting oligophenylenevinylene (OPV) moiety to an electron-conducting fullerene subunit.

The C60/3PV hybrid compound has been incorporated in a photovoltaic cell constructed by spin-casting the compound on a glass substrate covered with indium-tin oxide (ITO) and depositing an aluminum film on top. In such a device configuration, the compound is not only able to generate electrons and holes under light irradiation but it also provides pathways for their subsequent collection at opposite electrodes and a photocurrent is obtained.

We now report in more detail the preparative procedures used to obtain C60/3PV as well as the synthesis of the corresponding higher homologues C60/4PV.

We also discuss their electrochemical and excited-state properties in solution using the related OPV derivatives 3PV and 4PV and the fulleropyrrolidine FP as reference compounds.

Finally, their incorporation in photovoltaic devices is described.
It was recently shown that TEM images depend on the amount of defocus so that the wall thickness estimated in this way can vary by a factor of 2. It was also shown that the results of N$_2$ sorption isotherms strongly depend on the model that is used to analyze them. The calculation methods based on the Kelvin equation, in particular the method developed by Barrett, Joyner, and Halenda (BJH method), systematically underestimate pore diameters by at least 10%. In this work, we show how X-ray diffraction powder patterns can be quantitatively modeled to obtain lattice parameters, pore diameters, and wall thicknesses. We thus give very detailed new information about the silica organization in the walls of SBA-15 materials.

The mesoporous solids considered here, labeled SBA-15, are synthesized with triblock copolymers and have two-dimensional hexagonal symmetry. We used two neutral triblock copolymers (Pluronic (poly(ethylene oxide)-poly(propylene oxide)-poly(ethylene oxide) EO20PO70EO20 and EO18PO60EO18, BASF®, labeled P123 and P103, respectively) and tetraethyl orthosilicate (TEOS) as the silicate source. At room temperature, the poly(ethylene oxide) (PEO) chains are hydrophilic whereas the poly(propylene oxide) (PPO) chains tend to be hydrophobic, thus driving the formation of direct cylindrical micelles with the PEO chains on the outside. Consequently, a specific and interesting feature of the SBA-15 materials is that, as suggested by NMR, the PEO chains of the Pluronics may be deeply occluded within the silica walls and therefore the density of the walls may not be uniform (Figure 1). Then, the solids obtained upon calcination should display not only the mesoporosity (pore diameter >2 nm) due to the hexagonal arrangement of cylindrical aggregates but also the microporosity (pore diameter <2 nm) liberated by the PEO moieties. This would result in a complex structure in which a microporous silica “corona” surrounds the mesopores. Furthermore, hydrothermal treatments of as-synthesized SBA-15 materials in the reaction solution at different temperatures (80?140 °C) and for different times (11-72 h) are known to increase pore sizes and to decrease wall thicknesses.

To our knowledge, previous attempts to model the structure of organized mesoporous solids have only dealt with the MCM-41 materials obtained by templating with cationic surfactants. The most comprehensive one is the neutron scattering study of White et al., which shows that the mesopores of these materials are not as well-defined as commonly assumed. In contrast, there exists no detailed structural study of SBA-15 materials. We obtained here the dimensions and electron density contrasts of the different parts of the unit cell of these materials by modeling the intensities of their X-ray diffraction powder lines. This information proves that these solids should not be regarded either as an ideal hexagonal lattice of mesopores imbedded in a silica matrix. In particular, we demonstrate here the existence of the microporous corona and we monitor how it is affected by hydrothermal treatments.

Gold surfaces and anchor sulfur organic groups, such as thiols (RSH) and disulfides (RSSR) (where R is an alkyl chain), are the prototype components of supramolecular systems using self-assembled monolayers (SAMs) for the functionalization of extended surfaces and for the preparation of monolayer-protected clusters (MPCs). Given the wealth of potential technological applications, including corrosion inhibition, lithography, lubrication, catalysis, and molecular recognition, the interest in the underlying chemistry and physics has been steadily growing during the past decade. However, basic issues are still open, and the need to solve them has become urgent.

In particular the chemical processes accompanying the formation of the adsorbed species are largely unknown, the final interface and the nature of the headgroups have not been characterized to satisfaction, and the existence of dimers versus thiolates in the full-coverage regime of Au surfaces is still under debate. The adsorption on Au(111) of methanethiol and dimethyl disulfide (R = CH$_3$) at low coverage provides the simplest case, where the influence of van der Waals interactions is minimized. Also in this case a clear understanding is still missing, although experimental investigation of clean single-crystal surfaces under ultrahigh-vacuum conditions dates back to the beginning of the SAM proposal.

Partial attempts to establish the structure and energetics of these complex systems have been previously made which were however based on oversimplified models of the metal surface, consisting of small atomic clusters (with no or limited embedding), and on a limited set of calculations that ignored the possibility of the structural relaxation driven by adsorption processes. To make progress with the aid of computer simulations, we present a state-of-the-art computational approach which relies on a slab as the model of the unreconstructed Au(111) surface, is based on density-functional theory (DFT), and accounts for all structural modifications caused by adsorption processes. Indeed, Au(111) undergoes a 23 × reconstruction in which fcc domains alternate with hcp domains, separated by smaller regions with Au atoms on bridge positions. However, the effects of such a reconstruction on the electronic properties have previously been shown to be negligible.

We found that dissociation of the disulfide with formation of strongly bound thiolates is favored, in agreement with experimental evidence at low coverage, whereas thiolates resulting from S?H bond cleavage in thiols can coexist with the adsorbed “intact” species and become favored if accompanied by the formation of molecular hydrogen.
**Article#17:**

Recent experimental results obtained from a small group of molecules (ca. 1000) forming a nanopore showed a behavior that resembles a resonant tunneling diode (RTD) or a device with a negative differential molecular impedance. At a temperature of 60 K the current peaks at 2.1 V and reaches a peak current of 1 nA while the valley current is around 1 pA. The form of the peak is shown schematically in Figure 1. Also, as the temperature increases, the voltage at which the current is maximum decreases monotonically at a rate of 9 mV/K.

**Article#18:**

Self-assembly of metal-based supramolecules has attracted considerable interest because of the potential ability for selective inclusion and transportation of ions and molecules and the catalysis for specific chemical reactions.

**Article#19:**

Compounds in which unsaturated elemental carbon chains span two metals, LnMxLn’M’ (1; Scheme 1), constitute the most fundamental class of carbon-based molecular wires.
Such one-dimensional assemblies must be defined as composed only of sp-hybridized carbon.

They bear an obvious conceptual relationship to the classic two- and three-dimensional polymeric carbon allotropes, graphite and diamond, which are based upon sp2- and sp3-hybridized carbon. Importantly, all of these species must terminate with a non-carbon atom or some entity that differs from the repeat unit.

Hence, at very high chain lengths, I can be viewed as a genuine carbon allotrope.

§2

The polymeric sp carbon allotrope is often referred to as "carbyne", and numerous syntheses have been claimed. However, most samples are intractable or difficult to characterize.

As shown in Scheme 1, different valence structures are possible. One features alternating triple and single bonds (alkynyl or polyynediyl), with sp carbon termini bearing one endgroup X (II). Another consists solely of double bonds (cumulenic), with sp2 carbon termini bearing two endgroups X (III).

In the case of sp3 carbon termini with three endgroups X (IV), the bond order pattern of II is reversed. These distinctions become important when the Xn moieties are collectively replaced by a single redox-active metal capable of forming one, two, or three bonds to carbon.

§3

There is a fascinating older literature involving polyynes with bulky endgroups.

For example, tert-butyl-capped species Me3CC(CC)nMe3 (n = 4?7, 10, 12) have been prepared by classical techniques and characterized by UV?visible spectra and melting points.7 Similar R3Si(CC)nSiR3 species have been reported (n = 4?10, 12, 16), but the higher oligomers were generated in solution, characterized by UV?visible spectra, and not purified further.8

The physical properties of both series remain to be probed by modern methods.

More recently, Hirsch et al. have prepared mixtures of cyano-capped species NC(CC)nCN (n = 3?8) by graphite vapor deposition, and following HPLC separation carefully documented their properties.9 Lagow et al. have claimed similar syntheses of higher oligomers (n = 35?75) with cyano and trifluoromethyl endgroups.

However, systems with smaller endgroups are generally less stable and often explosive.

Cumulated species (III) also appear to be less stable.

Isolable examples are so far limited to C6 chains (n = 2).10

§4

Complexes of the type I have attracted the attention of numerous researchers from the standpoints of fundamental physical and chemical properties, materials attributes, molecular electronics, and catalysis, as summarized in reviews2 and previous full papers in this series.11,12

In this paper, we describe the systematic synthesis and detailed physical characterization of C6?C20 polyynes with the chiral rhenium endgroup (5-C2Me5)Re(NO)(PPh3) (Re).

This constitutes, together with the work of Hirsch et al. noted above, the first modern investigation of a homologous series of long polyynes.

We carefully document the effect of chain length upon the IR, Raman, 13C NMR, and UV?visible spectra, as well as redox potentials and thermal stabilities.

Portions of this work have been communicated.13,14

§5

Other important background details are as follows.

In a previous full paper,15 we described the isolation and physical properties of the three "consanguineous" or redox-related C4 complexes ReC4Ren+n(PF6) shown in Scheme 2.

This illustrates the interconversion of neutral polyynediyl and dicationic cumulenic valence structures, and the electrochemical generation of higher homologues is described below.

Related FeC4Fe, RuC4Ru, and MnC4Mn species, some of which can be even further oxidized, have been characterized by Lapinte, Bruce, and Berke.16

Some partially characterized C6 and C8 complexes of the type I had been reported at the outset of this work.17

Additional examples are now available,18,19 the most noteworthy of which is an FeC8Fe system of Lapinte that can be isolated in both neutral octatetrayneidyl and radical cation oxidation states.20

Outside of the title compounds, only two other complexes with longer chains (C12) have been described to date.13,21

Move Article#20:

§1

It is increasingly recognized that C?H??O hydrogen bonds play an important role in determining molecular conformation and crystal packing,2 in molecular recognition processes,3 in the stabilization of inclusion complexes,4 and possibly in the activity of biological macromolecules.5

Much of the evidence for C?H??O hydrogen bonding comes from the observation of close C??O contacts in crystal structures.

Although such interactions have been observed with a variety of possible C?H donors and oxygen acceptors, including amino acids,6 it is only recently that C?H??O hydrogen bonds have been observed in the crystal structures of proteins.

Surveys of high-resolution protein structures reveal the widespread occurrence of close C??O contacts.7,8

The majority of the reported contacts involve hydrogen atoms attached to the ?-carbons in the peptide backbone.

This is consistent with the polarized character of the C??O bond, adjacent to electron withdrawing N?H and CO groups.

The most frequently observed hydrogen bond acceptors are the carbonyl oxygens in adjacent peptide strands of ?-sheets (Figure 1).9

In these cases, one oxygen lone pair interacts with a C??O proton and the other oxygen lone pair interacts with an N?H proton.
The CH⋯O interaction has been recognized as playing a crucial role in molecular recognition. However, the strength of this interaction is not well-established. A recent computational study by Quiocho et al. suggested that CH⋯O hydrogen bonds are strong enough to stabilize complex structures. To investigate the nature of these interactions, we examined CH⋯O hydrogen bonding in a model system, the dimers of N,N,N,N'-tetramethylmalonamide and N,N,N,N'-tetramethylsuccinamide, which exhibit internal CH⋯O hydrogen bonds.

## §1

**Nonbonding interactions play a dominant role in many forefront areas of modern chemistry from molecular biology to material design.**

Detailed information of the interactions is essential for understanding biological processes and for simulations of materials.

The attraction between the C?H bond and the ? system is called CH⋯O interaction. It was first proposed by Nishio and co-workers to explain the preference of conformations in which bulky alkyl and phenyl groups had close contact.

During these two decades several experimental studies, which support the existence of the attraction, have been reported. The close contact was observed in stable conformations of a lot of molecules. Statistical analysis of the crystal structure database indicates that the short contact of the C?H bond and the ? system is observed in large numbers of organic crystals. The CH⋯O interaction is believed as a crucial driving force of crystal packing. The short contact was also observed in crystals of proteins. Aoyama et al. reported that the interaction is an important driving force of host-guest complexation of cyclic resorcinol derivatives. Stoddart et al. reported that self-organization of catenanes was governed by weak nonbonding interactions including the CH⋯O interaction. Quiocho et al. reported that crystal structures of sugar binding proteins show carbohydrate ligands sandwiched by aromatic side chains and the CH⋯O interaction is playing an important role in molecular recognition.

### Article#21:

**$\text{Cargiolo et al.}^{21}$**

Nonbonding interactions play a dominant role in many forefront areas of modern chemistry from molecular biology to material design. Detailed information of the interactions is essential for understanding biological processes and for simulations of materials. The attraction between the C?H bond and the ? system is called CH⋯O interaction.

It was first proposed by Nishio and co-workers to explain the preference of conformations in which bulky alkyl and phenyl groups had close contact. During these two decades several experimental studies, which support the existence of the attraction, have been reported. The close contact was observed in stable conformations of a lot of molecules. Statistical analysis of the crystal structure database indicates that the short contact of the C?H bond and the ? system is observed in large numbers of organic crystals. The CH⋯O interaction is believed as a crucial driving force of crystal packing. The short contact was also observed in crystals of proteins. Aoyama et al. reported that the interaction is an important driving force of host-guest complexation of cyclic resorcinol derivatives. Stoddart et al. reported that self-organization of catenanes was governed by weak nonbonding interactions including the CH⋯O interaction. Quiocho et al. reported that crystal structures of sugar binding proteins show carbohydrate ligands sandwiched by aromatic side chains and the CH⋯O interaction is playing an important role in molecular recognition.

### Appendix B

Move to Section: 1. This bifurcated hydrogen bonding pattern has also been observed in helical protein structures. A different bifurcated hydrogen bonding topology has been observed in a collagen triple helix in which both oxygen lone pairs contact C?H protons and one of the hydrogen atoms is shared between two oxygens (Figure 1).

2. Although it has been established that C?H⋯O interactions exist in protein structures, their role in determining structure is unclear.

3. It is generally assumed that the C?H⋯O hydrogen bond is much weaker than other interactions, e.g., N?H⋯O hydrogen bonds, between peptide strands.

4. Although the early calculations were impressive when they were reported, the basis sets used in these calculations (4-31G and MIDI) were too small to evaluate the weak attractive interaction energy quantitatively.

5. Nothing is known, either experimentally or theoretically, about the strength of this interaction in proteins.

6. It has been noted that C?H⋯O hydrogen bonds may be responsible for alleviating the unsaturated hydrogen bonding potential of buried polar groups with the suggestion that the total contribution of the C?H⋯O bonds may, at least in some proteins, play a significant role in stabilizing the folded structure.

7. It also has been suggested that C?H⋯O bonds probably should not be considered as a determinant of any particular folded structure, but that they more likely act to "fine-tune" local structure after the protein has adopted its final folded state.

8. A few theoretical calculations of model systems were carried out to evaluate the interaction energy of the CH⋯O interaction. Takagi et al. reported that main contributions to the attraction were electrostatic and charge-transfer terms from the energy decomposition analysis using the Kitaera/Morokuma scheme.

9. Therefore, charge transfer was believed as an important source for the attraction.

10. Although these early calculations were impressive when they were reported, the basis sets used in these calculations (4-31G and MIDI) were too small to evaluate the weak attractive interaction energy quantitatively.

11. Recently reported calculations of small hydrocarbon molecules show that the calculated interaction energies highly depend on basis set and electron correlation and that a large basis set including multiple polarization functions and appropriate electron correlation correction are necessary to accurately evaluate interaction energies.
Article#22:
§1
The factors contributing to the thermodynamic stability of the DNA double helix have been the focus of intense scrutiny for the past four decades.

§2
Both hydrogen bonding and base stacking are important as stabilizing noncovalent interactions in the double helical structure. Of these two, hydrogen bonding is perhaps the simpler and better understood interaction. Base stacking, although discussed at great length, is more complex and remains considerably less well understood.

§3
Despite this, it is clear that base stacking makes a strong contribution to stabilizing the helical structure of DNA and RNA. A number of studies in short RNA duplexes have made use of the "dangling end" effect, which occurs when a single unpaired base is added to the end of a duplex, stabilizing the helix by stacking on it. This method is highly useful since it separates the stacking interaction of a single base from other interactions involved in pairing (namely, stacking of the pairing partner and hydrogen bonding between the bases).

§4
In looped RNA or DNA structures, where hydrogen bonding is less extensive, it is likely that base stacking may be relatively even more important as a stabilizing interaction for helical structure.

§5
While stacking is by consensus an important noncovalent interaction in DNA and proteins, the nature of this interaction remains under debate. Theoretical studies have implicated several factors as potentially important in stabilizing the face-to-face base/base interactions. Among these are electrostatic (dipole/dipole and dipole/induced dipole) interactions, dispersion (momentary dipole/induced dipole) effects, and solvation effects. The electrostatic effects may depend on localized charges that exist at specific parts of a given heterocyclic base, or on electrostatic potentials that may differ between the faces and edges of the bases. Dispersion effects depend on the surface area of contact and on the polarizability of the two species. Finally, solvophobic and other solvent-driven effects will depend on the relative energies of solvation of bases when stacked and unstacked as well as the amount of surface area desolvated on stacking.

§6
While theoretical studies of base stacking exist in relative abundance, there have been fewer experimental studies of stacking in the context of DNA. Studies of nucleotide monomers, dimers, and related analogues have demonstrated stacking of unpaired bases in aqueous solution. Such studies have indicated that the relative stacking ability of the natural bases qualitatively goes in the order purine > pyrimidine > purine > pyrimidine. Since the addition of organic solvents destabilizes this interaction and nucleic acid duplexes in general, a solvophobic contribution to stacking has been implicated; however, this is not a classical entropy-driven hydrophobic interaction as seen in protein folding. DNA duplex formation is an enthalpy-driven process, researchers have concluded that any entropy-driven hydrophobic effects are hidden by unfavorable entropy of restricted bond rotations, and that electrostatic or van der Waals interactions (enthalpy-driven effects) may be more important in DNA than a solvent-induced interaction. On the whole, there is still not a unified picture as to the relative importance of solvophobic, electrostatic, and dispersive effects on stacking in water.

§7
In addition to studies with nucleic acids, a substantial number of simpler organic structures have been studied as models for aqueous stacking in the more complex nucleic acid and protein structures. For example, studies by Rebek and co-workers have shown that adenine analogues can be complexed by stacking with simple aromatic hydrocarbon groups; in one study it was found that increasing the size (surface area) of such a group increased binding significantly. Gellman et al. have studied bis-aromatic structures bridged by propylene units, suggesting that classical hydrophobic factors might not be as important in stacking interactions as was previously thought, although this has been debated. More recently the same group has studied a minimal "molecular balance" system to measure hydrophobic interactions quantitatively. Finally, work by other groups has stressed the importance of electrostatic factors in stacking interactions.

§8
Dougherty and co-workers have pointed out that the negative electrostatic potential in the center of the benzene ring face can be an important factor in its noncovalent interactions. Siegel has shown in elegant experiments that by adjusting this electrostatic potential with substituents one can affect the magnitude of interactions between closely aligned aromatic faces.

§9
Such model systems have led to valuable insights into the stacking question.
### Article#23:

#### §1

Important progress has been made over the last years to control free radical polymerizations and better understand the mechanisms involved.

Three main methods with their specificity and limitations have been shown to allow a controlled propagation of free radicals: either nitroxide-mediated radical polymerization (NRP), atom transfer radical polymerization (ATRP), or reversible addition/fragmentation chain transfer (RAFT). These methods can indeed be used to prepare well-defined polymeric samples from miscellaneous monomers.

#### §2

However, these three methods are not alike: compared to ATRP and RAFT, NRP appears until now of limited applicability; only two families of monomers, namely styrenes and alkyl acrylates, have actually been shown to polymerize under “living” controlled conditions in the presence of nitroxides.

Despite this, NRP still attracts much interest because it is based on potentially simpler systems than ATRP or RAFT and does not call for any added metal complexes or complicated syntheses.

Since the initial report by Georges and co-workers on the use of 2,2,6,6-tetramethyl-1-piperidinyloxy (TEMPO) to control the polymerization of styrene, many other contributions have been made to improve and expand NRP through the use of miscellaneous nitroxides.

For example, we showed in a previous report that N-tet-butyl-N-[1-diethylphosphono-2,2-dimethylpropyl)] nitroxide (DEPN) is not only efficient at controlling the radical polymerizations of styrene and of alkyl acrylates, affording samples of polystyrene (PS) and poly(n-butyl acrylate) (PBuA) with polydispersities lower than 1.1, but it also allows faster propagation than any other nitroxides.

#### §3

Thanks to the seminal investigations of Matyjaszewski, Georges, Fukuda, and Fischer and others, we now have a clear picture of the mechanism that is involved in polymerization mediated by TEMPO.

In this study, we explore the kinetics of polymerization of styrene and n-butyl acrylate with DEPN as chain growth moderator.

Mechanistically, this system should not be different from other controlled radical polymerizations that imply the reversible activation/deactivation of growing chains.

#### §4

Using the methodology described by Fukuda, we have determined the rate constant of activation (kd) of dormant PS-DEPN and PBuA-DEPN oligomeric adducts that were purposely prepared in separate experiments.

In situ ESR measurements have been concomitantly conducted in order to monitor the level of free DEPN formed as a function of time.

From the knowledge of the concentration of DEPN, the equilibrium constant (K = kd/krec) could be calculated; the rate constant of reversible deactivation (krec) was then easily deduced.

To check whether the rate constants measured and the mechanism proposed correspond to reality, we have used the PREDICI program and simulated the DEPN-mediated polymerization of styrene.

The main purpose of these simulations was to compare experimental data (conversion, [DEPN], etc.) with simulated ones and demonstrate that the rate constants measured are indeed consistent.

### Article#24:

#### §1

Prior to reports from these laboratories, a few Ni(II)-based catalysts were known to polymerize ethylene to linear polyethylene.

Due to rapid chain transfer, dimerization and oligomerization reactions of ethylene and ?-olefins were the most commonly reported features of late metal catalysts.

We have described a family of cationic palladium(II) and nickel(II) complexes bearing bulky aryl-substituted ?-diimine ligands which are capable of polymerizing ethylene, ?-olefins, and internal and cyclic olefins to high molar mass polymers.

Polymerizations of common, inexpensive monomers using these catalysts result in the formation of unique materials including highly branched, amorphous polyethylene, “chain-straightened” poly(?-olefins), and melt processable polycyclopentene.

In addition, the palladium-?-diimine catalysts are the first transition metal complexes reported which are capable of copolymerizing ethylene and ?-olefins with functionalized monomers such as alkyl acrylates and methyl vinyl ketone.

Following our initial reports of these catalysts, other groups have described similar results using ?-diimine-based systems.

Neutral, bidentate Ni(II) systems which incorporate bulky aryl-substituted imine units have also recently been shown to be quite effective for polymerization of ethylene to high molecular weight polymers.
In contrast to the extremely active, highly air-sensitive nickel analogues, \[^{\text{14}}\] the well-defined cationic palladium?alkyl species depicted in Figure 1 are more easily synthesized and handled and exhibit rates of reaction that make them well-suited for study using variable temperature NMR spectroscopic techniques.

On the basis of our initial NMR observations and analyses of polymer microstructures, we proposed the polymerization mechanism shown in Schemes 1A and 1B(i) (illustrated here for the case of ethylene polymerization).

The overall mechanism of polymerization consists of three main processes beyond initiation: (1) chain propagation, (2) metal migration along the polymer chain ("chain running"), and (3) chain transfer.

The first and third processes are common to all ethylene oligomerization and polymerization catalysts.

Metal migration, on the other hand, is insignificant for most catalytic systems \[^{\text{5,6}}\] but is a major reaction pathway for the Pd?N,N?diimine catalysts, thus making chain-running the most distinguishing feature of these systems.

For acyclic olefins, the resting states of the palladium catalysts have been observed by NMR spectroscopy to be alkyl?olefin complexes.

In the absence of excess olefin or other Lewis bases, the alkyl cations exist as stable ?-agostic species. \[^{\text{18,44}}\]

A ?-agostic alkyl species is observed to be the resting state for cyclopentene polymerizations. \[^{\text{18}}\]

Propagation and isomerization (Scheme 1A) are easily monitored by low-temperature NMR spectroscopy in this system.

Insertion of ethylene from the catalyst resting state \[^{\text{7}}\] produces the linear alkyl agostic species 8, which can be trapped by free ethylene.

Repetition of this cycle \[^{\text{7 ? 8 ? 7}}\] results in the growth of a linear polymer chain. Complex 8 can also ?-hydride eliminate and reinsert with opposite regiochemistry to produce the branched ?-agostic species 10.

This process may involve an alkyl?olefin intermediate, 9, though such a species has never been observed in this system.

The branched alkyl cation 10 can continue to "chain run" by ?-hydride elimination and reinsertion with opposite regiochemistry, producing longer branches.

Alternatively, 10 can be trapped with ethylene, and insertion produces a polymer chain containing a methyl branch along the backbone.

In the case of nickel, trapping and insertion are competitive with chain-running, and thus the extent of branching is dependent on ethylene pressure. \[^{\text{17 \& 18}}\]

For palladium systems, the total number of branches is nearly insensitive to ethylene pressure, but the number of branches-on-branches and the overall architecture of the polymer does vary with ethylene pressure. \[^{\text{5,6}}\]

In the case of ?-olefins, 2,1-insertion followed by metal migration to the terminal carbon can result in incorporation of ?-olefin in a 1,? (linear) fashion with the resultant polymer containing fewer branches than expected for simple 1,2 monomer enchainment, a phenomenon we term "chain-straightening". \[^{\text{18,20}}\]

Bulkly aryl-substituted ?-diimine complexes slow the rate of chain transfer relative to propagation, resulting in the formation of high polymer.

This rate retardation clearly results from substituting the ortho positions of the aryl rings with bulky substituents which then project into the axial sites of the square planar complexes. \[^{\text{21}}\]

The precise mode of chain transfer is not yet completely understood.

Initially we proposed that chain transfer occurred by associative displacement of the unsaturated chain from an olefin hydride complex (Scheme 1B(i)), a process slowed by axial bulk.

Calculations by Ziegler suggest chain transfer may occur from the alkyl olefin resting state by direct ?-hydride transfer to monomer (Scheme 1B(ii)). \[^{\text{18,20}}\]

We describe in this paper mechanistic investigations of olefin polymerizations catalyzed by the well-defined cationic palladium?alkyl complexes depicted in Figure 1.

These investigations include NMR spectroscopic characterization of catalyst resting states, measurement of insertion barriers and the effects of substituents on these barriers, determination of the mode and energetics of palladium migration, as well as the nature and relative stabilities of key ?-agostic alkyl complexes, and measurement of associative exchange rates in alkyl?olefin complexes which relate to the chain transfer process.

These experimental studies are complemented by recent theoretical calculations from other groups concerning various mechanistic details of these catalyst systems. \[^{\text{18,44}}\]

Article#25:

DNA has been well-known for over 50 years as the genetic material of living systems.

Our understanding of the chemical basis for its genetic role \[^{\text{1}}\] has led to the explosive growth of molecular biology since the 1950s and its application in molecular biotechnology since the 1970s.

Recently, two new fields have aimed to exploit the chemical properties of DNA:

These are DNA nanotechnology \[^{\text{5,6}}\] and DNA-based computation. \[^{\text{5}}\]

The goals of DNA nanotechnology include the construction of nanoscale objects \[^{\text{5}}\] and devices \[^{\text{5}}\] from DNA, as well as the self-assembly of periodic arrays. \[^{\text{5,6}}\]

DNA nanotechnology takes advantage of the fact that the intermolecular interactions of DNA are diverse, highly specific, and readily programmed through Watson?Crick complementarity.

This complementarity can be used to design the spontaneous assembly of single strands into double helices, branched junctions, \[^{\text{4}}\] or more complex motifs. \[^{\text{7}}\]

In addition, the Watson?Crick complementary associations of single-stranded overhangs ("sticky ends") can be used in the same way, to
DNA nanotechnology and some of the approaches to DNA-based computing are dependent on unusual motifs of DNA. The key feature of the most useful motifs is that they contrast with normal DNA in that their helix axes are not strictly linear, but are arranged to flank branch points. The most prominent unusual DNA motif is the immobile DNA branched junction, a stable analogue of the Holliday intermediate in genetic recombination. Branched molecules have been constructed that contain three, four, five, or six helices flanking a branch point. Other unusual motifs have included knots, antijunctions and mesojunctions, and double crossover (DX) molecules.

One powerful technique for constructing DNA nanostructures and for performing DNA-based computation is to use the self-assembly of DNA tiles. Wang tiles are theoretical constructs that contain colored edges, and they assemble so that edges with the same colors abut, much like dominoes. It is straightforward to imagine placing four-arm DNA branched junctions at the centers of square tiles, with their helix axes directed normal to the edges. Equating complementary sticky ends with the colored edges reduces the Wang tiles to a concrete self-assembling molecular form, using the medium of DNA. Whereas the assembly of Wang tiles can emulate the operation of a Turing machine, assembling branched junctions in this fashion can lead to meaningful computation; it has been shown that self-assembled arrays of DNA molecules can be used in principle as cellular automata.

Thus, experimentally, DNA molecules, referred to here as DNA tiles, can be designed to self-assemble into arrays such as those of Wang tiles through the associations of specific sticky ends. Previous demonstrations of large scale two-dimensional DNA self-assemblies can be regarded as experimental verification of the feasibility of such tilings. However, the self-assembly of DNA tilings used for computation has not yet been demonstrated experimentally. The success of the self-assembly approach will depend critically on advances in the design of the individual DNA tiles.

Like the Holliday junction, the DNA double crossover molecule (DX) is an analogue of intermediates in genetic recombination, both meiotic recombination and recombination mediated by double strand breaks. The DX molecule consists of two double helical domains joined twice at crossover points; another way to think of the DX molecule is that it is formed from two four-arm branched molecules that have been ligated at adjacent arms. There are five unique motifs to the DX molecule, illustrated in Figure 1a.

The three parallel DX motifs, DPE, DPOW, and DPON, are relevant to biological processes, but they are not stable when the separations between their crossovers are two turns or fewer. By contrast, the antiparallel DX molecules, DAE and DAO are stable and well behaved, both in solution, and when analyzed on nondenaturing polyacrylamide gels.

The DAE molecule contains an even number of double helical half-turns between crossover points, whereas the separation in the DAO molecule is an odd number of half-turns. One consequence of this difference is that the DAE molecule consists of four strands of DNA, while the DAO molecule contains an additional cyclic molecule between the crossover points; as a practical point, it is difficult to seal this cyclic strand. Recently, we have analyzed DAE molecules in ligation-closure experiments where the ligation is catalyzed by T4 DNA ligase.

It is easiest to analyze these experiments if the product of ligation contains a reporter strand, which is a strand whose fate reflects that of the complex. Figure 1b shows that a ligated oligomer of DAE molecules contains a reporter strand, in contrast to DAO molecules.

This is the reason that DAE molecules were analyzed before DAO molecules. The striking result of this analysis is that DAE molecules appear to be quite stiff, in contrast to conventional branched molecules. This finding suggested that DX molecules would be useful for the construction of large DNA arrays:

In fact, they have been used recently to construct large arrangements of DNA both in one dimension and in two dimensions, although not yet analyzed in the same way, DAO molecules have also proved to be rigid enough to form two-dimensional arrays.

Winfree has demonstrated that DX molecules can also be used as cellular automata.

Both DNA nanotechnology and DNA-based computing would benefit from a greater diversity of DNA motifs, particularly those based on branched molecules.
<table>
<thead>
<tr>
<th>Move</th>
<th>Article#26:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perylene derivatives are promising compounds for application in electronic devices as molecular semiconductors.</td>
</tr>
<tr>
<td>1</td>
<td>Most organic conducting materials can be described as p-type semiconductors, in which holes in the valence band are the majority charge carriers.</td>
</tr>
<tr>
<td>1</td>
<td>In contrast, in N,N-dialkyl-3,4,9,10-perylenetetracarboxyldimides the majority charge carriers are electrons in the conduction band, and this material is thus classified as an n-type semiconductor.</td>
</tr>
<tr>
<td>1</td>
<td>Such materials are potentially useful as the electron-accepting materials in all organic photovoltaic solar cells, as was reported in 1986 for the first solar cell based on a phthalocyanine and a perylenetetracarboxylic derivative.</td>
</tr>
<tr>
<td>1</td>
<td>This type of solar cell is based on a p/n junction in which the phthalocyanine is the p-type material (electron donor) and the perylene derivative is the n-type material (electron acceptor).</td>
</tr>
<tr>
<td>2</td>
<td>Since then, much research has been done on this kind of solar cell, but the energy conversion efficiencies are still limited to about 1%, as was, e.g., obtained for p/n junctions of zinc phthalocyanines (50 nm ZnPc) and N,N-dimethyl-3,4,9,10-perylenetetracarboxyldimides (25/50 nm PTCDI).</td>
</tr>
<tr>
<td>2</td>
<td>Ways to improve the efficiency of the p/n junction include an increase of the exciton diffusion length and an improvement of the transport of charge carriers in the organic material.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
<tr>
<td>2</td>
<td>Optimization of these aspects allows for the use of thicker light-absorbing layers, which has the effect that a higher percentage of the incident sunlight can be absorbed.</td>
</tr>
</tbody>
</table>
§2

Many different phthalocyanines\(^\text{1}\) (and other electron-donating materials\(^\text{2}\)) have already been synthesized and characterized, but the perylene part has received much less attention.

It has been shown that increased ordering of the perylene derivatives in the presence of the vapor of an organic solvent leads to an increase in exciton diffusion length.\(^\text{3}\)

For an N,N'-bis(phenethylimide)-3,4,9,10-perylenebicarboximide film, which has been annealed by exposure to methylene chloride vapor for 3 h, an exciton transfer length of 2.5 ± 0.5 \(\text{Å}\) has been reported.\(^\text{4}\)

This value is the longest yet reported for a perylene diimide derivative.

The increased ordering of the perylene derivative film after vapor annealing also stabilizes charge-separated states, which slows charge recombination.\(^\text{5}\)

2 Therefore, to increase the efficiency of organic p/n junction solar cells, it is useful to design highly ordered perylene derivatives.

Liquid crystals possess the ability to self-organize spontaneously into highly ordered structures.

Due to the spontaneous ordering of the molecules in the liquid crystalline phase, very highly ordered films can be obtained after the liquid crystalline film is cooled to a crystalline phase.

2 Therefore, the use of liquid crystalline perylene derivatives in organic p/n junctions is a promising method to increase the efficiency of such a solar cell.

§3

2 So far, only a few liquid crystalline perylene derivatives have been synthesized.\(^\text{6}\)

3 In this paper the detailed phase behavior is discussed of N,N'-dialkyl-3,4,9,10-perylenetetracarboxyldiimides (Figure 1, with \(n = 7\) (1), 12 (2), and 18 (3)).

3 The phase transitions were determined using differential scanning calorimetry (DSC) and polarized optical microscopy (POM).

3 The three-dimensional ordering of these perylene derivatives is examined with X-ray diffraction.\(^\text{7}\)

3 The cell parameters found by the X-ray diffraction measurements have been used in molecular mechanics simulations to find the lowest energetically favorable packing of the molecules in the liquid crystalline phases.

3 Together with UV/vis spectroscopy studies on evaporated thin films of the perylene derivative, these measurements and simulations provide a detailed picture of the average ordering of these molecules.

§4

3 To determine the usefulness of these perylene derivatives as semiconducting materials, e.g., in organic solar cells, the mobility of charge in N,N'-dioctadecyl-3,4,9,10-perylenecarboximide (3) has been measured using the pulse-radiolysis time-resolved microwave conductivity (PR-TRMC) technique.\(^\text{8}\)

3 PR-TRMC measurements yield the sum of the positive and negative charge carrier mobilities, \(\mu_{\text{TRMC}} = \mu^+ + \mu^\text{e}\).

3 Electrons have, however, been found on the majority charge carriers in this type of material, and would therefore be expected to contribute predominantly to \(\mu_{\text{TRMC}}\). Because of the nanosecond time scale of monitoring and the gigahertz electric field frequencies used, the mobilities obtained are trap-free values associated with well-organized domains within the material.\(^\text{9}\)

3 The value of \(\mu_{\text{TRMC}}\) should therefore be close to the maximum mobility that could be obtained in a dc device with a monodomain of that particular material between the electrodes.

3 Very good agreement between mobility values determined by PR-TRMC and by dc time-of-flight (TOF) measurements has in fact been found for the liquid crystalline phases of the mesomorphic discotic material hexa(hexylthio)triphenylene, which forms well-organized, interelectrode layers.\(^\text{10}\)

Article#27:

§1

1 Studies of nucleic acids using molecular dynamics simulations have shown considerable progress during the past five years.\(^\text{11}\)

1 With the use of particle-mesh Ewald to treat long-range electrostatics, stable trajectories of DNA and RNA molecules up to several nanoseconds have been obtained.\(^\text{12}\)

1 Details of the structural and dynamical properties of various nucleic acids, ranging from conventional DNA and RNA duplexes to DNA:RNA hybrids, and RNA hairpin loops, can be elucidated through these MD studies.

1 To accurately represent the environment, such simulations generally are carried out in the presence of explicit solvent molecules.

1 The box of solvent molecules must be large enough to minimize electrostatic interactions between periodic images of the solute, generally leaving at least 10 Å between each edge of the box and the closest solute atom.

1 This results in a system of around 3000 water molecules for a 10-base pair duplex DNA.\(^\text{13}\)

1 In addition to increasing calculation time, complications also arise from the need to fully equilibrate these molecules and any counterions in the system.

1 Even with today’s powerful computational facilities, these simulations can be lengthy and costly.

§3

1 The electrostatic effects of a solvent of high dielectric, such as water, can be approximated by a continuum electrostatics model.\(^\text{14}\)

1 This model has been extensively studied, mostly through numerical solutions of electrostatic equation in a multiple dielectric model.

1 The generalized Born (GB) model\(^\text{15}\) provides an approximate solution to the solute?solvent electrostatic polarization term (?Gpol).

1 This model is often able to reproduce the solvation energies given by the Poisson?Boltzmann (PB) continuum solvent model for a variety of biomolecules, without the costly computations of the numerical solutions to Poisson's equation.\(^\text{16}\)

1 Furthermore, an extension to the GB model to account for salt effects in the linearized Debye?Hückel approximation has been shown to give good agreement with PB results.\(^\text{17}\)

1 Although many early applications of this model considered the energetics of a relatively small number of conformations (such as snapshots of
Article#28:
§1
There has been a growing interest in the design and synthesis of polymer nanostructures with ever-increasing degrees of complexity and control over the composition and structure.

§2
We have chosen to focus on shell cross-linked (SCK) polymer micelles to investigate the effect of the cross-linked membrane-like shell on the transport properties of guests and to from the core, while maintaining a mobile core domain.

§3
Additionally, since stabilization is provided by binding together of the shell, the polymer chains that constitute the core can be degraded and extracted to leave hallow nanoencage networks.

§4
The cross-linking, which has been achieved by reactions performed within the core domain or throughout the shell layer, introduces covalent stabilization for the polymer micelles to reinfore the micellar architecture and result in high stability toward dilution and shear forces.

§5
As a mimic of globular proteins, synthetic nanoparticles having core/shell morphology have been prepared by the self-assembly of block copolymers, to afford polymer micelles, followed by intramicellar cross-linking.

§6
Shell cross-linked micelles with hydrophilic core and zwitterionic properties have also been prepared recently. These materials are essentially single molecules of nanometer-scale dimensions.

§7
Therefore, they exhibit unique and robust character, which avoid the morphological changes that can occur for their supramolecular micellar precursors under environmental variations, such as changes in ionic strength, solvent system, and pH.

§8
Additionally, variation over the core and shell composition and dimensions has been demonstrated, and the stability achieved through covalent cross-linking suggests that these nanostructures can serve in a capacity similar to that of micelles, for example, in the transport and delivery of guest molecules under particularly demanding conditions.

Article#29:
§1
The treatment of human diseases through the application of nucleotide-based drugs such as DNA has the potential to revolutionize the medical field.

§2
While viral vectors are currently the most efficient gene transfer agents, concerns surrounding the overall safety of viral vectors, which include the potential for unsolicited immune responses, have resulted in parallel efforts to develop nonviral alternatives.

§3
Cationic polymers have been widely used as transfection vectors due to the facility with which they condense and protect negatively charged strands of DNA.

§4
As a result, the choice of cationic polymer for a gene transfer application generally requires a tradeoff between transfection efficiency and short- or long-term cytotoxicity.
Additionally, the long-term biocompatibility of these polymers remains an important issue for use in therapeutic applications in vivo, since several of these polymers are not readily biodegradable.\(^1\) 

For use in vivo, new cationic polymers should be designed incorporating hydrolyzable moieties such that the polymers readily degrade into nontoxic byproducts.\(^1\) 

To develop safe alternatives to existing polymeric vectors and other functionalized biomaterials, we and others have pursued the synthesis of degradable polyesters bearing cationic side chains.\(^1\) 

Examples of such polymers include poly(l-lactide-co-l-lysine), poly(serine ester), poly(4-hydroxy-l-proline ester), and, more recently, poly[2-(4-aminobutyl)-l-glycolic acid].\(^1\) 

Poly(4-hydroxy-l-proline ester) and poly[2-(4-aminobutyl)-l-glycolic acid] were recently demonstrated to condense plasmid DNA through electrostatic interactions and mediate gene transfer.\(^1\) 

Generally, these new polymers are less-toxic than poly(l-lysine) and PEI and they degrade into nontoxic metabolites.\(^1\) 

It is clear from these investigations that the rational design of amine-containing polyesters can be a productive route to the development of safe, effective transfection vectors.\(^1\) 

Unfortunately, however, present syntheses of these polymers require either the independent preparation of specialized monomers or the use of stoichiometric amounts of expensive coupling reagents.\(^1\) 

Additionally, the amine functionalities in the monomers must be protected prior to polymerization, necessitating additional postpolymerization deprotection steps before the polymers can be used as transfection agents.\(^1\) 

We sought to develop a complementary strategy for the synthesis of amine-containing polyesters that would address several of these synthetic drawbacks.\(^1\) 

While degradability, reduced cytotoxicity, and an ability to complex DNA remained fundamental design criteria, we also desired an approach that would yield a class of polymers that were structurally different from previously studied polymeric vectors so that structure/activity relationships could be further explored.\(^1\) 

To date, most cationic polymers synthesized and investigated for transfection efficacy have contained pendant amines in the polymer side chains.\(^1\) 

Herein, we report a strategy for the preparation of poly(2-aminosters) containing tertiary amines in their backbones based on the conjugate addition of bis(secondary amine) monomers to diacylate esters (eq 1).\(^1\) 

These polymers interact electrostatically with plasmid DNA at physiological pH and assemble it into nanometer-scale polymer/DNA complexes.\(^1\) 

The polymers themselves are nontoxic and they degrade into nontoxic small molecule byproducts.\(^1\) 

---

**Move**

**Article#30:**

1. Although there is a large body of literature concerning transition metal doped semiconductors, also called diluted magnetic semiconductors (DMS), and there has been a flurry of activity investigating semiconductor quantum dots (QDs), it is only recently that there have been reports describing research occurring at the intersection of these two fields.\(^1\) 

2. Of the work detailing the synthesis of DMS QDs, the majority rely on photoluminescence (PL) spectroscopy to determine if the materials have been successfully doped: below band gap energy emission at the wavelength corresponding to the bulk material is the primary litmus test for doped nanocrystals.\(^1\) 

3. However, this emission only implies that electronic coupling exists between the nanocrystal and the Mn excited states.\(^1\) 

4. Although it is then usually inferred that the Mn dopant must be inside the QD lattice, the structural evidence supporting these claims is often minimal.\(^1\) 

5. An exception is work reported by Kanatzidis and co-workers, which presents nanocrystalline Cd1-xMnxSe powder X-ray diffraction patterns displaying with textbook clarity the linear dependence of the lattice constant with x.\(^1\) 

6. Although their method is powerful because it allows x to be varied continuously from 0 to 1, the material is too large (~200 Å) to exhibit quantum confinement.\(^1\) 

**§2**

1. A recent report from Ladizhansky and co-workers concludes that the dopant in their CdMnS samples, prepared by the popular coprecipitation method, is either in the solvent/ligand matrix or bound to the surface of the particles.\(^1\) 

2. In light of these results, any thorough characterization of doped QDs must include convincing structural evidence, or numerous purification steps and control experiments, that determines the location of the dopant atoms from among a number of possibilities: trapped in the matrix containing the sample, bound to the protecting surface ligands, adsorbed onto the surface of the particles, doped randomly within the inorganic lattice, etc.\(^1\) 

**§3**

1. By incorporating manganese into the QD one can investigate the effects of a paramagnetic dopant, and the effective magnetic field it generates, on the semiconductor properties.\(^1\) 

2. Here we are particularly concerned with how the dopant affects the characteristics of the quantum dot optical properties.\(^1\) 

3. The paramagnetic impurity provides a means of coupling the optical and magnetic properties of these materials via the sp?d exchange interaction.\(^1\) 

4. This strong interaction between the carriers of the semiconductor and the spin of the dopant is well-known in bulk DMS; it is at the origin of the unique properties of these materials, including giant Faraday rotation effects near the band edge.\(^1\) 

5. The preparation of doped QDs should allow the study of the effects of three-dimensional quantum confinement on well-known bulk DMS properties.\(^1\) 

6. A report by Awschalom and co-workers of changes in the magnetic behavior of thin CdMnTe films as the layer thickness is reduced from
86 to 18 Å suggests that quantum confinement should affect the manganese/manganese interaction (the d-d exchange) as well.

Manganese-doped nanocrystals are likely to display properties which have not been observed in either bulk DMS or undoped semiconductor QDs.

In this paper we describe the use of the single source manganese- and chalcogen-containing organometallic complex Mn2(?-SeMe)2(CO)8 to prepare Cd1-xMnxSe QDs of various sizes.

Simpler precursors containing only manganese were ineffective dopants when incorporated into the trioctylphosphine oxide (TOPO) method for making CdSe QDs.

Extensive surface ligand exchange was performed to ensure that our quantitative measurements detected only manganese ions incorporated into the lattice, rather than species loosely bound to the nanocrystal surface.

For Cd1-xMnxSe QDs, there is good agreement between manganese concentrations determined by wavelength dispersive X-ray spectroscopy (WDS) and results from electron paramagnetic resonance (EPR) measurements.

Magic angle spinning solid-state 113Cd nuclear magnetic resonance (NMR) spectra of doped QDs display the paramagnetic shift and decreased longitudinal relaxation times reported for the corresponding bulk material.

We study the effect on the optical properties via fluorescence line narrowing (FLN) and photoluminescence excitation (PLE) experiments.

By fitting the resulting spectra we extract emission and absorption line shapes where inhomogeneities associated with the residual size distribution have been significantly reduced.

The extracted emission line shapes display trends which had previously been seen in undoped nanocrystals under the influence of an external magnetic field.

The corresponding extracted absorption line shapes of the doped QDs can be viewed as excitation spectra in the presence of a magnetic field.

Such an experiment has not been performed before this point, and it reveals new structure in the absorption line shape.

This new feature can be explained in the context of a dark exciton model of the band edge luminescence.

The convergence of results from doped QDs and undoped QDs in a magnetic field confirms that transition metal doping provides a powerful and simple means of performing high magnetic field experiments without an external magnet.
Appendix C: Overview of Statistical Analysis

Table 4. Correlation between Reference age and Vector based similarity.

<table>
<thead>
<tr>
<th>Year of reference</th>
<th>Vector similarity for highly cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of reference</td>
<td>Pearson Correlation</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Reference</td>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>849</td>
<td>,297**</td>
</tr>
<tr>
<td>,000</td>
<td>N</td>
</tr>
<tr>
<td>849</td>
<td>849</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

References Grouped in Sections

Fig. 6. Boxplot diagram showing vector based similarity for highly cited references across sections. Outliers and extreme values are labeled with year.
Table 5. Descriptive data for vector based similarity and age of highly cited references across sections.

<table>
<thead>
<tr>
<th>Section</th>
<th>Vector similarity for highly cited</th>
<th>Year of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>Mean</td>
<td>0.94398975</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>478</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.0915566850</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.065155620</td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td>Mean</td>
<td>0.028624668</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.0444473910</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.011308482</td>
</tr>
<tr>
<td><strong>Results &amp; discussion</strong></td>
<td>Mean</td>
<td>0.050615345</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.0653891367</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.023108558</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>Mean</td>
<td>0.017036097</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.0240073764</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.001499697</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>Mean</td>
<td>0.071727102</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>849</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.0831904432</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.041259358</td>
</tr>
</tbody>
</table>
Fig. 7. Histogram over vector based similarity of highly-cited references tiered in ten equal-sized groups (1 lowest, 10 highest), and colorcoded by section.

Table 6. Crosstabulation showing distribution of references in tiered similarity groups according to section and age.

<table>
<thead>
<tr>
<th>Year of reference</th>
<th>Section</th>
<th>Tiered Similarity Groups</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td>Introduction</td>
<td></td>
<td>8</td>
<td>4</td>
<td>17</td>
<td>26</td>
<td>28</td>
<td>28</td>
<td>37</td>
<td>37</td>
<td>48</td>
<td>59</td>
<td>292</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Results &amp; discussion</td>
<td></td>
<td>15</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>11</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>27</td>
<td>17</td>
<td>36</td>
<td>36</td>
<td>45</td>
<td>41</td>
<td>55</td>
<td>54</td>
<td>66</td>
<td>72</td>
<td>438</td>
</tr>
<tr>
<td>before 1996</td>
<td>Section</td>
<td></td>
<td>9</td>
<td>14</td>
<td>18</td>
<td>29</td>
<td>24</td>
<td>25</td>
<td>17</td>
<td>21</td>
<td>18</td>
<td>11</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td></td>
<td>18</td>
<td>34</td>
<td>16</td>
<td>11</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td></td>
<td>29</td>
<td>20</td>
<td>26</td>
<td>9</td>
<td>8</td>
<td>14</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>Results &amp; discussion</td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
<td>58</td>
<td>68</td>
<td>60</td>
<td>49</td>
<td>40</td>
<td>44</td>
<td>30</td>
<td>31</td>
<td>19</td>
<td>12</td>
<td>411</td>
</tr>
</tbody>
</table>
Table 7. Chi-square tests for the significance of section on the similarity score of the highly cited references.

<table>
<thead>
<tr>
<th>Old reference</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td>Pearson Chi-Square</td>
<td>52,880&lt;sup&gt;a&lt;/sup&gt;</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>48,185</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>22,988</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>438</td>
<td></td>
</tr>
<tr>
<td>before 1996</td>
<td>Pearson Chi-Square</td>
<td>116,485&lt;sup&gt;b&lt;/sup&gt;</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>124,683</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>54,095</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>411</td>
<td></td>
</tr>
</tbody>
</table>

a. 21 cells (52.5%) have expected count less than 5. The minimum expected count is .08.

b. 13 cells (32.5%) have expected count less than 5. The minimum expected count is .09.

Moves Coded by Sentence

Fig. 8. Boxplot diagram showing vector based similarity for highly cited references across sentence-level moves. Outliers and extreme values are labeled with year.
Table 8. Descriptive data for vector based similarity and age of highly cited references across sentence-level moves.

<table>
<thead>
<tr>
<th>Move coded by sentence</th>
<th>Vector similarity for highly cited</th>
<th>Year of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mean</td>
<td>.099394279</td>
<td>1994.20</td>
</tr>
<tr>
<td>N</td>
<td>422</td>
<td>760</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>.0940505269</td>
<td>8.279</td>
</tr>
<tr>
<td>Median</td>
<td>.070590713</td>
<td>1997.00</td>
</tr>
<tr>
<td>2 Mean</td>
<td>.078711935</td>
<td>1991.67</td>
</tr>
<tr>
<td>N</td>
<td>63</td>
<td>134</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>.0748119421</td>
<td>9.780</td>
</tr>
<tr>
<td>Median</td>
<td>.056497428</td>
<td>1995.00</td>
</tr>
<tr>
<td>3 Mean</td>
<td>.091773994</td>
<td>1994.18</td>
</tr>
<tr>
<td>N</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>.0955448658</td>
<td>7.236</td>
</tr>
<tr>
<td>Median</td>
<td>.079358653</td>
<td>1996.00</td>
</tr>
<tr>
<td>Total Mean</td>
<td>.096559695</td>
<td>1993.83</td>
</tr>
<tr>
<td>N</td>
<td>500</td>
<td>934</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>.0919894279</td>
<td>8.507</td>
</tr>
<tr>
<td>Median</td>
<td>.068183983</td>
<td>1997.00</td>
</tr>
</tbody>
</table>
**Fig. 9.** Histogram over vector based similarity for highly cited references tiered in ten groups (1 lowest, 10 highest), and color-coded by sentence-level moves.

**Table 9.** Crosstabulation showing distribution of references in tiered similarity groups according to sentence-level move and age.

<table>
<thead>
<tr>
<th>Year of reference</th>
<th>Tiered Similarity Groups</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td>Move coded by sentence</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>14</td>
<td>25</td>
<td>22</td>
<td>25</td>
<td>34</td>
<td>34</td>
<td>45</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>8</td>
<td>4</td>
<td>17</td>
<td>26</td>
<td>28</td>
<td>29</td>
<td>40</td>
<td>41</td>
<td>52</td>
<td>65</td>
<td>155</td>
</tr>
<tr>
<td>before 1996</td>
<td>Move coded by sentence</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>16</td>
<td>25</td>
<td>19</td>
<td>20</td>
<td>14</td>
<td>21</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>9</td>
<td>14</td>
<td>19</td>
<td>29</td>
<td>24</td>
<td>25</td>
<td>18</td>
<td>23</td>
<td>18</td>
<td>11</td>
<td>190</td>
</tr>
</tbody>
</table>
Table 10. Chi-square tests for the significance of sentence-level move on the similarity score of highly cited references.

<table>
<thead>
<tr>
<th>Old reference</th>
<th>Pearson Chi-Square</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td>19,044^a</td>
<td>18</td>
<td>.389</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>19,128</td>
<td>18</td>
<td>.384</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>.549</td>
<td>1</td>
<td>.459</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>before 1996</td>
<td>18,736^b</td>
<td>18</td>
<td>.408</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>17,662</td>
<td>18</td>
<td>.478</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>.768</td>
<td>1</td>
<td>.381</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>190</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. 19 cells (63.3%) have expected count less than 5. The minimum expected count is .09.
b. 20 cells (66.7%) have expected count less than 5. The minimum expected count is .38.

Moves Coded by Paragraph

Fig. 10. Boxplot diagram showing vector based similarity for highly cited references across paragraph-level moves. Outliers and extreme values are labeled with year.
Table 11. Descriptive data for vector based similarity and age of highly cited references across paragraph-level moves.

<table>
<thead>
<tr>
<th>Move coded by paragraph</th>
<th>Vector similarity for highly cited</th>
<th>Year of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean : 0.091025099</td>
<td>1993.76</td>
</tr>
<tr>
<td></td>
<td>N : 241</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation : 0.0953679760</td>
<td>8,115</td>
</tr>
<tr>
<td></td>
<td>Median : 0.056203529</td>
<td>1996.00</td>
</tr>
<tr>
<td>2</td>
<td>Mean : 0.086697199</td>
<td>1992.88</td>
</tr>
<tr>
<td></td>
<td>N : 177</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation : 0.0766990055</td>
<td>10,102</td>
</tr>
<tr>
<td></td>
<td>Median : 0.069344429</td>
<td>1997.00</td>
</tr>
<tr>
<td>3</td>
<td>Mean : 0.134114570</td>
<td>1995.71</td>
</tr>
<tr>
<td></td>
<td>N : 82</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation : 0.1032880612</td>
<td>5,534</td>
</tr>
<tr>
<td></td>
<td>Median : 0.119001679</td>
<td>1997.00</td>
</tr>
<tr>
<td>Total</td>
<td>Mean : 0.096559695</td>
<td>1993.83</td>
</tr>
<tr>
<td></td>
<td>N : 500</td>
<td>934</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation : 0.0919894279</td>
<td>8,507</td>
</tr>
<tr>
<td></td>
<td>Median : 0.068183983</td>
<td>1997.00</td>
</tr>
</tbody>
</table>
Fig. 11. Histogram over vector based similarity for highly cited references tiered in ten groups (1 lowest, 10 highest), and color-coded by paragraph-level moves.

Table 12. Crosstabulation showing distribution of highly cited references in tiered similarity groups according to age and paragraph-level move.

<table>
<thead>
<tr>
<th>Year of reference</th>
<th>Tiered Similarity Groups</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995 Move coded by paragraph</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>16</td>
<td>12</td>
<td>21</td>
<td>31</td>
<td>131</td>
</tr>
<tr>
<td>Move coded by paragraph</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>22</td>
<td>14</td>
<td>18</td>
<td>122</td>
</tr>
<tr>
<td>Move coded by paragraph</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>17</td>
<td>16</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>4</td>
<td>17</td>
<td>26</td>
<td>28</td>
<td>29</td>
<td>40</td>
<td>41</td>
<td>52</td>
<td>65</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>before 1996 Move coded by paragraph</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>13</td>
<td>19</td>
<td>13</td>
<td>14</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>110</td>
</tr>
<tr>
<td>Move coded by paragraph</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>Move coded by paragraph</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>14</td>
<td>19</td>
<td>29</td>
<td>24</td>
<td>25</td>
<td>18</td>
<td>23</td>
<td>18</td>
<td>11</td>
<td>190</td>
<td></td>
</tr>
</tbody>
</table>
Table 13. Chi-square tests for the significance of paragraph-level move on the similarity score of highly cited references.

<table>
<thead>
<tr>
<th>Old reference</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>29,357(^a)</td>
<td>18</td>
<td>.044</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>29,249</td>
<td>18</td>
<td>.045</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>4,342</td>
<td>1</td>
<td>.037</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>before 1996</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>18,728(^b)</td>
<td>18</td>
<td>.409</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>20,636</td>
<td>18</td>
<td>.298</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>6,811</td>
<td>1</td>
<td>.009</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>190</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. 8 cells (26.7%) have expected count less than 5. The minimum expected count is 1.74.
b. 13 cells (43.3%) have expected count less than 5. The minimum expected count is 1.18.

Moves Coded Algorithmically

![Boxplot diagram showing vector based similarity for highly cited references across moves coded algorithmically. Outliers and extreme values are labeled with year.](image)

Fig. 12. Boxplot diagram showing vector based similarity for highly cited references across moves coded algorithmically. Outliers and extreme values are labeled with year.
Table 14. Descriptive data for vector based similarity and age of highly cited references across moves coded algorithmically.

<table>
<thead>
<tr>
<th>Move coded algorithmically</th>
<th>Vector similarity for highly cited</th>
<th>Year of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean</td>
<td>1993,93</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>288</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0,0845875774</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0,060029054</td>
</tr>
<tr>
<td>2</td>
<td>Mean</td>
<td>1993,04</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0,0799093267</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0,077643695</td>
</tr>
<tr>
<td>3</td>
<td>Mean</td>
<td>1994,47</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0,1202326346</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0,102270331</td>
</tr>
<tr>
<td>Total</td>
<td>Mean</td>
<td>1993,75</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>468</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0,0889245659</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0,064516912</td>
</tr>
</tbody>
</table>
Fig. 13. Histogram over vector based similarity for highly cited references tiered in ten categories (1 lowest, 10 highest), and color-coded by algorithmic moves.

Table 15. Crosstabulation showing distribution of highly cited references in tiered similarity groups according to age and algorithmically coded move.

<table>
<thead>
<tr>
<th>Year of reference</th>
<th>Tiered Similarity Groups</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>after 1995 Move coded algorithmically</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>before 1996 Move coded algorithmically</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 16. Chi-square tests for the significance of algorithmically coded move on the similarity score of highly cited references.

<table>
<thead>
<tr>
<th>Old reference</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 1995</td>
<td>Pearson Chi-Square</td>
<td>44,003&lt;sup&gt;a&lt;/sup&gt;</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>46,006</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>13,271</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>284</td>
<td>-</td>
</tr>
<tr>
<td>before 1996</td>
<td>Pearson Chi-Square</td>
<td>26,256&lt;sup&gt;b&lt;/sup&gt;</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>28,385</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>2,242</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>184</td>
<td>-</td>
</tr>
</tbody>
</table>

a. 14 cells (46.7%) have expected count less than 5. The minimum expected count is .39.
b. 17 cells (56.7%) have expected count less than 5. The minimum expected count is 1.22.