Automatic Analysis of Author Judgment in Scientific Articles Based on Semantic Annotation

Marc Bertin and **Iana Atanassova** and **Jean-Pierre Descles**
Paris-Sorbonne University
28 rue Serpente 75006 Paris

Abstract

In this paper we describe how the annotation methodology adopted in our approach allows us to explain the organization of indexed references in scientific research articles. We identify the semantic values of author judgments in the text segments containing indexed references. We use an automated semantic annotation platform to annotate our corpora. Exploiting this result, we obtain a representation of the annotation distribution on different scales. Finally, we present two evaluations of the annotation.

1. Introduction

Science evaluation is based mainly on scientific publications, which are based on bibliographic citations. The study of bibliographic citations is a key element in the comprehension and elaboration of bibliometric indicators. In this article we discuss the problem of author judgment in scientific research articles. Moed (Moed 2005) suggests in his book "Citation Analysis in Research Evaluation" the necessity of a new approach. The efforts in the development of science evaluation methods should be directed towards qualitative citation analysis "through contextual and cognitive-relational analysis". To give a solution to some of these problems, we have developped a methodology based on the automatic semantic annotation of scientific research articles. We aim to identify relations between authors, and more specifically what is related to author judgments by their peers. The question which arises is the following: Are there any traces in the scientific literature of the authors' motivation to cite other work that can be automatically identified?

We know that there exist different motivations for citation. In 1964, and then in 1977, Garfield (Garfield 1977) identified fifteen different reasons for citation, such as for example: paying homage to pioneers, giving credit for related work, identifying methodology, providing background reading, correcting a work, criticizing previous work, etc. In 1982, Small (Small 1982) for his part made five such distinctions and proposed the following classification: Refuted, Noted Only, Reviewed, Applied, Supported. These categories can be considered respectively as: negative, perfunctory, compared, used and substantiated.

The problem can also be stated as follows: How can we identify the relations based on judgment between authors? Our approach should give some elements of response to this problem. We will explain it in detail in the next section.

In section 2 we show how this approach can be carried out using semantic annotation of scientific research articles. We describe our methodology in three steps: identification of the indicator, use of linguistic clues and automatic annotation of corpora. For the last step, we use a platform for the semantic annotation of corpora of scientific articles, based on linguistic and computational methods. In section 3 we describe our corpus and explain the reason for its constitution, then we present and comment the results we have obtained by this method. In section 4 we present the evaluation of the approach in two stages: an evaluation of the indicators and an evaluation of the obtained annotations. Finally, we discuss the relevance of this work and the perspectives for this kind of applications.

2. Method

In this section we present and explain our method for the analysis of author judgment. More particularly, we are interested in the elaboration of a method which permits to annotate automatically corpora of scientific articles in view to identify and categorize author judgments in texts.

We want to verify if the act of citation can be categorized by examining the text of a scientific article. Is it true that indexed references are only simple pointers to the bibliography or can we use them to localize the textual segments that will be candidates for annotation? We suppose that author judgments are expressed in scientific articles in the textual space close to the indexed references. The value of this hypothesis is developed in the discussion section.

Firstly, we will consider only the text segments which have a semantic annotation, and then, we will analyze the distribution of the annotations. We identify the indexed references, that we consider as indicators, by using finite state automata. After the identification of these indicators, we can use a research space (to the left and to the right of the indexed reference) for the application of contextual exploration rules. This approach is based on the Contextual Exploration Method (Desclés 2006): the semantic value of the

indicator is determined by the presence or absence of linguistic clues. By using an initial corpus, we have created linguistic ressources that will be used by an automatic annotation plateform.

Here we will describe in detail the three main stages of this method. In this part, we maintain a clear distinction between the Finite State Automata, the linguistic rules and the annotation engine.

2.1 Finite State Automata

The first step of our approach is to locate the textual segments in which we will most probably find the judgment of the author on another author. To do this we have adopted the hypothesis that the author judgments are localized in the textual space close to an indexed reference.

To identify the indexed references in texts we have constructed finite state automata. It must be noted that there exist different norms for bibliographic references that can be used, namely the norms ISO-690 and ISO 690-2, which are the international standards from the International Organization for Standardization, as well as the French norms AFNOR NF Z 44-005 and AFNOR NF Z 44-005-2. The second version of the international and French norms are related to numeric documents. However, a finite state automata based only on these norms is not sufficient to carry out the text processing on a large scale. In practice the norms are not rigorously applied by authors of scientific texts. That is why in order to create the finite state automata different corrections had to be made to take into consideration the different customs in writing indexed references. In this way we determine the finite state automata which identify the textual segments containing indexed references. This methodology is proposed in (Bertin et al. 2006).

The implementation of this methodology is given by a list of regular expressions that we have constituted in order to identify the different indexed reference forms. Figure 1 presents a part of our list of regular expressions.

```
<marqueur no="401">(\[)[5][t][A-Z][a-z]+[A-Z][a-z]+[A-Z][a-z]+[0-9][3](\])</marqueur>
<marqueur no="402">\([A-Z][a-z]+(,( )?[A-Z][a-z]+)[0,5](,)?( )?(et ([A-Z][a-z]+))(,)?( )?[0-9][4]\)</marqueur>
<marqueur no="403">\(voir par exemple [A-Z][a-z]+ et [A-Z][a-z]+, [0-9][4]\)</marqueur>
<marqueur no="404">\(([A-Z][a-z]+(,( )?)?(([0-9][4])(;),)?( )?)?)[0,5](et( )?)[A-Z][a-z]+(,|;)( )?[0-9][4]\)</marqueur>
<marqueur no="405">\(Adapté de ([A-Z][a-z]+(,( )?)?(([0-9][4])(;),)?( )?)?)[0,5](et( )?)[A-Z][a-z]+(,|;)( )?[0-9][4]\)</marqueur>
<marqueur no="406">\(([A-Z][\']?[A-Z][a-z]+( )et( )([A-Z][\']?[A-Z][a-z]+,( )?[0-9][4])\)</marqueur>
<marqueur no="407">\(([A-Z][a-z]+(, )?(, et )?){0,5}[A-Z][a-z]+(, et )?(, [0-9][4]), [0-9][4]\)</marqueur>
<marqueur no="408">\(Adapté de ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9][4])\)</marqueur>
<marqueur no="409">\(D'après ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9][4])\)</marqueur>
<marqueur no="410">\(([A-Z][\']?[A-Z][a-z]+(( )et( ))?(, )?){1,6}(et ([A-Z][\']?[A-Z][a-z]+)(, )[0-9][4]\)</marqueur>
```

Figure 1: List of Regular Expressions

The different variations in the form of the indexed references had to be taken into consideration in order to obtain this list. For example, there are various phenomena such as, for example, the presence of commentaries in the indexed references, which makes them more difficult to identify. Similarly, the variety of sources (monographes, reviews, conferences, whether or not the citation refers to a chapter, page, etc.) and the different origins of the sources play an important role in the implementation of the automata.

All these reasons lead to the fact that the finite state automata necessary for the identification of the indexed references are of considerable complexity.

Table 1 presents a classification of the different types of indexed references that can be found in scientific articles.

	[1]	[AUT-08]	[Author, 2008]	Einstein
Nature	Indexed reference			
			Named entity	
Identification	Regular expression			
			Named entity recognition	
Norms	ISO 690 and ISO 690-2			
Epistemology	Frontier knowledge			
	Core knowledge			
Out of context comprehension	None	Resercher from the domain	Researcher	General
>> Growing complexity for the identification >>				

Table 1: Reference Classification

The difficulties of this type of approach are mainly related to the problems of named entity recognition as well as the problems of name transcriptions. We will not develop these issues in this article but we have to take into consideration this limitation. For the next step, regular expressions are not enough for the identification of author judgements, so we will develop the next stage, which consists in using linguistic clues.

2.2 Contextual Exploration

By studying an initial corpus we determine the linguistic resources that will be used by the annotation platform. The platform uses the notions of indicator, linguistic clues and contextual exploration rules.

The linguistic units identified by the automata are used as indicators in the framework of the Contextual Exploration Method. This is a decision-making procedure, presented in the form of a set of rules and linguistic markers that trigger the application of the rules.

The method consists firstly in the identification in the text of linguistic units, called indicators, that carry the semantic meaning of the categories for annotation. The presence of an indicator in a text segment is necessary for the application of the Contextual Exploration rules. The localization of complementary linguistic clues which are co-present in the context of the indicator permits the attribution of a semantic annotation to the segment. By establishing an hierarchy between the indicators and complementary linguistic clues, the Contextual Exploration Method defines an inference process: the presence of an indicator of a semantic category in a given segment corresponds to the hypothesis that this segment belongs to the category. The application of the Contextual Exploration rules permits to affirm or negate this hypothesis, or in some cases refine the category, and eventually annotate the segment.

The rules identify, in the left or right context of the indicator, linguistic clues that permit to take a decision, i.e. the annotation of the segment with the corresponding category. Both the indicators and the clues can be words or expressions, regular expressions, or the absence of an expression.

The main categories are organised in a semantic map, that has been published in (Bertin et al. 2006) and is fully operational in this implementation. The categories in this seman-

tic map will be used for the classification of the scientific texts. The implementation is presented on figure 2.

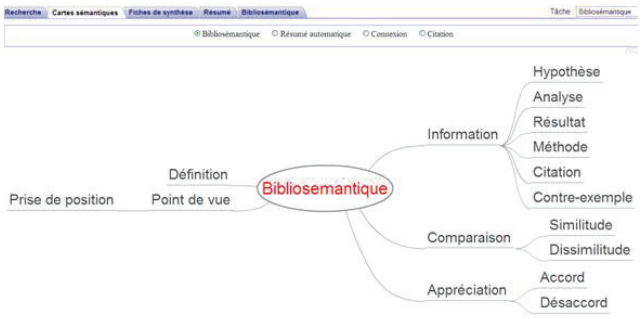

Figure 2: Implementation of the semantic map

2.3 Semantic Annotation Engine

For the annotation we use the platform for automatic semantic annotation called EXCOM. This platform has been developed in Java by (Alrahabi and Desclés 2008) and is available online on: http://www.excom.fr. It takes as input the linguistic ressources and carries out the segmentation and automatic annotation of texts. The system uses the UTF-8 encoding which permits the processing of different languages. The output is in the XML file format, according to the DocBook standard, where the annotations are presented as attributes to the text segments.

2.4 Semantic Annotation Interface

The results obtained from the automatic annotation are imported into an SQL database, designed for this purpose. This database contains, among others, the segments and the semantic annotations, as well as the links between the annotated segments and the initial texts. Our interface for information retrieval and bibliosemantic analysis is based on this database, in which the annotation results are exploited.

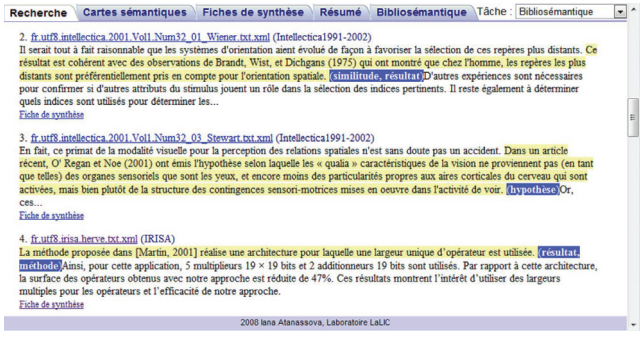

Figure 3: Annotation interface

The aim of this interface is to provide the possibilities for the analysis of the obtained annotations, as well as the graphic representation of the annotated segments. The interface (Atanassova et al. 2008) is developped in PHP/MySQL. It is common to other projects for semantic annotation and provides a number of functionalities, among which:

- information retrieval according to the annotated categories (figure 3);
- graphic visualization and dynamic management of the semantic map;
- construction of categorized syntheses of texts and corpora;
- visualization of graphs for the bibliosemantic analysis.

3. Annotation experiment

3.1 Corpus

Due to the lack of a standardized evaluation corpus of French scientific research articles, we have constructed our own corpus, the content of which is described in table 2. Our corpora consist of papers in the domains of linguistics and computer science. The initial corpus which was used for the creation of the linguistic resources consists of several papers selected randomly from this larger corpus.

Corpus	Language	Coverage	Format
CALS	fr	33 textes	Pdf
LaLIC	fr	8 textes	Doc/Pdf
ALSIC	fr/eng	1998-2007	Pdf/html
TALN	fr/eng	1999-2005	Pdf
Intellectica	fr/eng	1991-2002	Pdf
IRISA	fr/eng	1984-2006	Pdf/ps
PhD Theses	fr	6 theses	Pdf

Table 2: Corpus

3.2 Annotation Results

The results presented in this article are issued of a small part of our corpora which contains about 200 scientific papers and 6 PhD theses. The results of the automatic annotation with the linguistic resources show that in a scientific article the annotation distribution is not homogeneous. Our analysis shows that annotations are not evenly distributed in the text. The obtained data indicates that it is useful to study the annotation distribution, because these data can provide new possibilities for the comprehension of the citation phenomenon.

We can therefore envisage to make a categorization of scientific texts according to the annotations.

We can propose an analysis of this phenomenon on different scales:

- a part of a text: for example a chapter;
- a text, such as an article or a PhD thesis;
- a set of texts, or a corpus.

By extension, each document containing indexed references can be analysed by this methodology. The graph on figure 4 presents the results for several articles of the Intellectica corpus. The categories of the semantic map are represented on the horizontal axis and the vertical axis represents the position of the segment in the text. In our implementation, the textual segments correspond to sentences in the text.

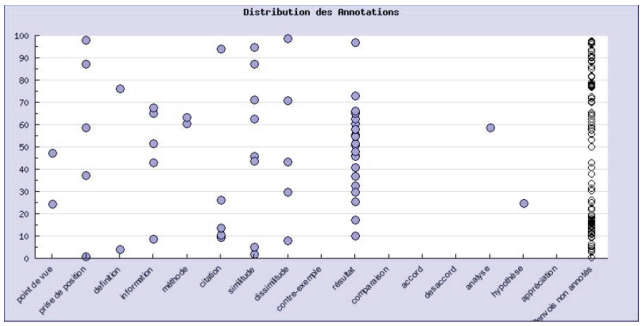

Figure 4: Annotation distribution of articles from the Intellectica corpus

The figure 4 shows that in this corpus the indexed references represent a wide variety of categories. The annotations are relatively well distributed, which shows that the different categories are well present and play a role in the scientific argumentation. This is not always the case. There exist other texts for which this in not valid. For example, we can consider a chapter of a PhD thesis in our corpus, for which we have obtained the representation on figure 5.

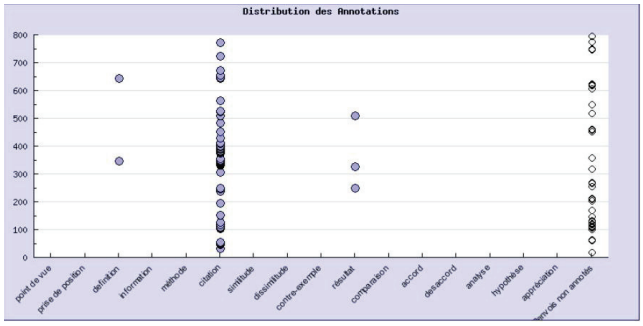

Figure 5: Annotation distribution of a PhD thesis chapter

We can see on this graph that there is one dominant category which is the most present in this text: the category of citation. Few segments are annotated with other categories, and these are namely results and definitions. This shows that this chapter is rather informative with less argumentation and author judgments.

Given these results, we can consider a categorization of scientific texts or corpora, based on the annotated segments. For example the graph on figure 6 gives an image of the Intellectica corpus analysed above, on which the numbers of segments annotated with the main categories of the semantic map are presented on the axes. This allows us to identify the dominant categories in the texts.

In this graph it is useful to note that the category of information is the most important. In fact this category contains other sub-categories which are frequently used in scientific articles, such as result, method, analysis, etc. However, the categories of comparison and opinion are also very well represented. In general, it is in these two categories that we would find author judgments.

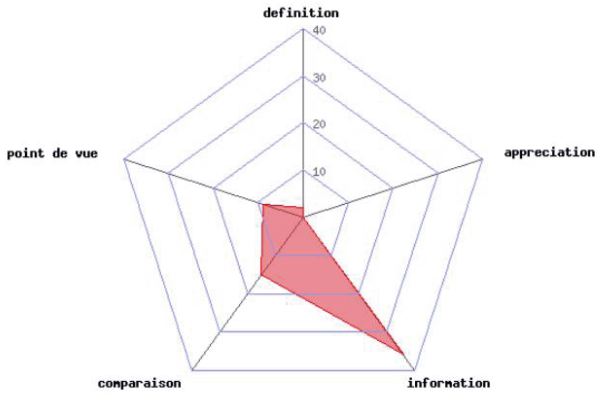

Figure 6: Categorization: Intellectica corpus

4. Evaluations

The evaluations we have carried out aim to establish the validity of our approach, i.e. finite state automata that we have constructed and the linguistic resources used to obtain the semantic annotations.

We have carried out a first evaluation concerning only the identification of the indicators. This is important because this evaluation can show whether the set of textual segments related to our hypothesis has been well identified. As all the semantic annotation is triggered by these indicators, thier identification is crucial for the next stages of analysis. More precisely, the Contextual Exploration rules are triggered by indicators in order to annotate the textual segments. The silence caused by the non-identification of indexed references will lead to silence in the stage of annotation. For this evaluation we have used the measures of precision and recall.

The second evaluation consisted in testing the validity of the automatic annotations carried out by the platform. The result that this evaluation should provide is more qualitative. We have adopted a methodology based on the Kappa coefficient, which is presented further on together with a discussion on this type of evaluation.

4.1 Evaluation I: Precision and recall measures

The first evaluation consists of measuring the accuracy of the retained indicators, or indexed references, which have been identified automatically by the finite state automata. We can proceed to the calculation of the traditional precision/recall measures(Salton and McGill 1979; Rijsbergen 1979) which determine the capacity of the system to correctly identify the textual segments containing an indicator.

The scores of precision and recall are defined by using the following quantities:

- TP (true positives): the number of segments that are correctly identified by the system;
- FN (false negatives): the number of segments that correspond to the automate but are not identified by the system;
- FP (false positives): the number of segments that are incorrectly identified by the system.

Then, for the recall and precision measures, we have:

$$R(M) = \frac{TP}{TP + FP}\ , P(M) = \frac{TP}{FP + FN}\ . \quad (1)$$

The results we have obtained by this evaluation are presented in table 3.

Recall	Precision
91,09%	98,91%

Table 3: Evaluation of the Indicators

We consider that these results are satisfactory. It must be noted that there is very little noise which means that almost all of the identified indicators are valid. On the other hand, the value of the recall is also very high. The several percents of indexed references not identified by the system are due to the various orthography rules for the names in different languages, as well as the presence of commentaries in the indexed reference itself.

Considering the classification of the indexed references, presented on table 1, we have made the same evaluation once more, but this time taking into consideration the named entities in the text. As the finite state automata that we have defined are not designed to identify named entities, the value of the recall for this evaluation has diminished significantly (table 4).

Recall	Precision
67,15%	98,91%

Table 4: Evaluation of the Indicators and Named entities

This result is due to the fact that our automata is designed to identify the indexed references and not the named entities in the scientific articles. The use of named entities in scientific articles can be explained as follows: after the first mention of the indexed reference in the text, certain authors prefer to use the named entity rather than the indexed reference. A second reason is that there exist authors who are cited directly without mentioning them in the bibliography, because their work is very well known and has been integrated in the common knowledge of the domain.

4.2 Evaluation II : Cohen's weighted Kappa

The problem we have to consider is how to evaluate the semantic annotation which is by definition qualitative in nature. The difficulty in the evaluation of NLP systems arises with the fact that although the form of the desired system output can be specified, the quality or the correctness of this output is difficult to define formally. As Popescu-Belis(Popescu-Belis 1999) says:

> There are however NLP tasks to which our evaluation model does not seem suited. These tasks often generate natural language texts as (part of) their result. Thus, the correct answers cannot be determined in advance. For instance, automatic summarization yields a summary whose quality can only be assessed by human judges [...]. Proposals for summarization evaluation attempt to measure a person's capacity to fulfill some tasks (e.g., document classification) using only the summaries. Likewise, for automatic translation, the utility of the translated texts is a measure of their quality. It seems thus that evaluation of such tasks cannot be automated.

In particular, in the case of semantic annotations, the definition of the correctness of the output is often given by examples or in reference to human faculties (linguistic competence, categorization capacity or other knowledge) that have not yet been completely formalized. Thus, the difficulty in defining the correct system output of the semantic annotation presents a problem for the automatic evaluation of such systems. The necessity of a formal framework arises for this kind of evaluations. This variability is punished in a number of scientific domains, where it is often necessary to evaluate or improve the coherence of the informations having the same nature and applied to the same object. This is even more important for the semantic annotations.

We can consider the agreement between a number of human evaluators on the annotation of a textual segment as a relevant measure in this case. The test Kappa (K) proposed by Cohen (Cohen 1960) provides a method to measure numerically the agreement between two or more observators or methods in the case when the judgments are qualitative in nature. This test consists in carrying out a session of "concordance" between the judges in order to evaluate the rates of agreement between them by the Kappa coefficient and also to study the cases of disagreement in view to improve the results. The agreement between the judges is defined as the conformity of two or more informations concerning the same object.

We have adopted this method for the second stage of our evaluation. In order to carry out the test, we have constituted a base of annotated text segments and these segments have been evaluated independently by two human judges. The judges had to classify the segments into two categories: correct and incorrect. In the table 5 we present the results of this evaluation.

		Judge A		
Judge B	Reponses	correct	incorrect	**Total**
	correct	77	10	87
	incorrect	6	7	13
	Total	83	17	100

Table 5: Evaluation results

The work of Landis and Koch (Landis and Koch 1977) propose a classification of the agreement as a function of the value of the Kappa coefficient. For example, a coefficient larger than 0,81 shows that the agreement is excellent. Between 0,80 and 0,61 it is good and between 0,20 and 0 the agreement is bad. The value of the Kappa coefficient that we have obtained by the evaluation according to table 5 is $K = 0,83$. We could therefore conclude that the agreement between the judges is excellent.

5. Discussion

For French scientific research articles, this study shows the existence of the phenomena of using author judgment through literature. Our approach permits the annotation of the segments containing indexed references. Our experience shows that in a number of cases the textual segments containing indexed references cannot be annotated. This is due to several reasons:

1. The indexed reference can play only the role of a pointer to a bibliography item and does not have any semantic meaning. In this case there is no author judgment explicitly expressed in the text.
2. The author judgment can be expressed in another text segment than the one containing the indexed reference. As our starting hypothesis was that the author judgment can be found in the same text segment as the indexed reference, we cannot identify these author judgments for the moment.

We consider that these two limitations do not invalidate our approach, but give us way to improve it. In fact, our aim is to identify and classify author judgments expressed in texts, if they exist and if they are explicitly stated. We want to annotate our corpora qualitatively and not quantitatively.

6. Conclusion and perspectives

The approach that we propose, namely the identification of textual segments based on indexed references, is relevant for the bibliosemantic analysis. The discussion has shown that we could improve some of the results, especially by considering larger research spaces. Given the obtained results, we could make some reflections on the existing methods for science evaluation. We have to ask ourselves the question whether we could continue to consider the bibliography as the basic unit for bibliometrics tools. The existence of different categories and the possibility to annotate them automatically by a semantic annotation platform gives the opportunity to raise the discussion on the validity of the existing methodologies for scientific evaluation and could give ground to study more in detail the nature of bibliometrics indicators.

References

Alrahabi, M., and Desclés, J.-P. 2008. Automatic annotation of direct reported speech in arabic and french, according to semantic map of enunciative modalities. *6th International Conference on Natural Language Processing, GoTAL 2008*.

Atanassova, I.; Desclés, J.-P.; le Priol, F.; and Franchi, A. 2008. La plate-forme excom comme outil automatique d'annotations smantiques des textes pour la catgorisation d'informations sur le web. *Internet: besoin de communiquer autrement, University St. Clement Of Ohride, Sofia, Bulgaria 2008*.

Bertin, M.; Desclés, J.-P.; Djioua, B.; and Krushkov, Y. 2006. Automatic annotation in text for bibliometrics use. In *FLAIRS 2006, Florida*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20:27–46.

Desclés, J.-P. 2006. Contextual exploration processing for discourse automatic annotations of texts. *FLAIRS 2006, Florida. Invited Speaker*.

Garfield, E. 1977. Can citation indexing be automated ? *Essay of an Information Scientist* 1.

Landis, J., and Koch, G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–174.

Moed, H. 2005. *Citation Analysis in Research Evaluation*. Springer.

Popescu-Belis, A. 1999. Evaluation of natural language processing systems: a model for coherence verification of quality measures. *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment. ELSE Project LE4-8340 (Evaluation in Language and Speech Engineering)*.

Rijsbergen, C. V. 1979. *Information Retrieval*. Butterworth-Heinemann Newton, MA, USA.

Salton, G., and McGill, M. J. 1979. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Small, H. 1982. Citation context analysis. *B. Dervin and M. Voigt (Eds.), Progress in communication sciences* 3:287–310.