RHECITAS: citation analysis of French humanities articles

Ludovic Tanguy[1], Fanny Lalleman[1], Claire François[2], Philippe Muller[3] and Patrick Séguéla[4]
1: CLLE-ERSS: CNRS and University of Toulouse
2: INIST: CNRS
3: IRIT: CNRS and University of Toulouse
4: Synapse Développement
tanguy@univ-tlse2.fr, fanny.lalleman@etu.univ-tlse2.fr,
claire.francois@inist.fr,muller@irit.fr, patrick.seguela@synapse-fr.com

Abstract
The RHECITAS project aims at the analysis of citations in French Humanities and Social Sciences articles using natural language processing techniques. It is based on a corpus of online articles, through the aid of natural language processing tools. The project is funded by TGE-ADONIS (CNRS, French National Research Centre). Although very little research, either theoretical and technical, has been made on such data (most approaches focusing on science publications written in English), we developed two different tools that can automatically *a)* identify the more important items in a list of references, based on a number of linguistic cues, and *b)* extract relevant terms associated to a reference. These results show a new angle on citation analysis, both from a linguistic point of view and for practical applications.

1 Introduction

The work presented here consists in the first steps in the RHECITAS project, which aims at the operational study of citations in French Humanities and Social Sciences (hereafter HSS) online publications. The RHECITAS project is funded by the French National Centre for Scientific Research (CNRS), and more precisely by the TGE-ADONIS subdivision. TGE-ADONIS' main goal is to promote digital humanities in France through the development of generic tools and resources. In this context, the long-term objective of the RHECITAS project is to enrich online publications portals in order to improve their accessibility and diffusion.
Citation analysis is now a major field in information and documentation science, and has led to important achievements such as citation indexes, and to cultural evolutions in the academic world, through the development of bibliometrics (or quantitative analysis of publications). However, most of the efforts so far have been limited to scientific publications in English: other academic disciplines and other languages have still to be taken into account. To our knowledge, no previous research work has addressed the analysis of French HSS citations.
As will be shown in the following, the lack of availability of ready-to-use data led us to design and apply new approaches to citation analysis. More precisely, we had to keep away from quantitative and network-oriented approaches and to focus on citation contexts, and thus to apply corpus linguistics methods to investigate some of the dimensions of citation behaviour.
We defined two more specific objectives. The first one is the automatic distinction between "important" citations and "background" or perfunctory ones. This distinction, although less ambitious than existing citation function typologies, is more appropriate to humanities articles and to a multi-disciplinary corpus. The second objective is the extraction of specific information about what the author "borrowed" from a reference. Our approach is based on specific patterns and cue-phrases, specifically developed for this application, and fully automated using NLP techniques. In this regard, we follow the work of Teufel et al (2006) on the identification of citation function.

For the first objective, we designed an "integration scale" for citations, measuring the embedding of the citation marker in the author's discourse (Swales 1990). Features used for this task are the location and syntactic function of the citation, as well as a number of more specific cues. The first results allow us to automatically provide a short list of the most important references for a given article. This task is described in section 4.
The second objective is to provide the reader with more specific information for each reference selected by the previous method. We presently focus on the extraction of key terms or concepts associated with a reference. We based our approach on specific patterns, such as quotation marks and definitions of terms (eg "X, which Y calls Z"), as shown in section 5.
Technically, *Rhecitas* consists of a fully operational tool that performs the automatic annotation (in XML format) and the subsequent analysis. A module is responsible for harvesting online articles and extracting references and citation contexts. The GATE platform is used for analysing citation contexts, using the French syntactic parser Cordial Analyseur, and a set of local grammars and lexical resources for the extraction of the linguistic features. The data and tools used in this project are described in section 3.
Some issues and aspects of citation analysis are presented in section 6, in which we outline further developments.

2 Overview of citation analysis approaches

Although “citation analysis” is a commonly used term, its meaning covers at least three different types of approaches, which we describe in details in the following sections.

2.1 Bibliometrics

The most common approach to citation analysis in the broad academic world is best known as bibliometrics (or scientometrics). A scholarly field in itself, bibliometrics aims at the investigation of research activities through the analysis of publications. This field has been defined by Eugene Garfield in the 1960’s (see for example Garfield 1962) and has considerably grown since that time. The main tool for bibliometrics is the *citation index*, i.e. a database containing, for a large number of scientific publications, the list of cited references.
Several such databases have been built and are made available to the academic community, mainly Gardfield’s own ISI and its different indexes (Web of Science, Web of Knowledge), Elsevier’s Scopus, and more recently Google Scholar.
Such databases can be queried in a number of interesting ways, enabling the answers to questions such as:

- “How many publications cite article X/author Y/journal Z?”
- “Which articles/authors/journals are frequently cited together?”

The first set of questions is used in research evaluation, when a high number of citations are assumed to be correlated to the quality of an author, article or journal. This area has led to the definition of a number of complex measures, such as the impact factor, h-index and so on. Although this purely numerical aspect of research evaluation is highly controversial in the academic world, it has been made widely available, for example through online databases and applications such as Anne-Wil Harzing’s “*Publish or Perish*” tool (`www.harzing.com`).
The second set of questions, less controversial, addresses the identification of *research fronts*, i.e. new efforts in scientific activities that can be matched with emerging clusters in citation relations, when a subset of authors are found to address similar questions and as such are frequently citing each others. This usage of citation indexes is of interest to documentation specialists and to people interested in the monitoring of science evolutions (Moed 2005).

Whatever the complexity of these approaches, the data on which they are based is limited to simple connexions between publications, without taking any qualitative aspects into account. More precisely, all citations are estimated to be of equal importance and validity, whatever the author's reason for citing another's work.

2.2 Citation classification

To address the previous point, another approach to citation analysis has focused on the qualitative characterization of a given citation. A number of documentation specialists have investigated the many different reasons why a given article is cited, and some have even tried to define contextual clues to identify these reasons. Although current developments in this area have not yet led to usable tools or data, the need to make the distinction between citations is commonly raised, at least to provide a counterpoint to the purely numerical aspects of bibliometrics (Moed 2005). The means to answer the question "*Why does X cite Y?*" fall into one of the following two possibilities. One can first simply ask the authors' opinion about their own behaviour, through specific and focused surveys, which of course require a lot of resources and cannot be applied on a large scale. The other possibility is to investigate the citation context, i.e. the part of the citing article's text where the citation occurs. This second method has the advantage of being more objective, and can be automated through natural language processing methods.

To achieve a citation classification, the first step is to establish a list of categories, or typology, of citations functions. This has been addressed since Garfield's pioneer work, and a large number of such typologies have been established (see Bornmann and Daniel 2008 for an excellent overview).

One of the main problems is the dependency of these typologies on specific research fields: most of them have been established through focused surveys, as shown in the extracts from three different typologies (figure 1). Garfield's typology aims at covering the general academic world, but has an understandable bias towards sciences; Krampen and Montada's categories are specific to clinical psychology, while the more recent proposition by Teufel et al. has been established on a study of computational linguistics articles. Although these typologies show significant variations (both in specific categories and in the underlying motivations), a few higher level regularities can be extracted: perfunctory citations (paying homage, simple mention, neutral, etc.), criticism of cited work, use of data/method taken from cited work, etc.

Amongst the controversies about bibliometrics is the fact that negative citations (where the citing author criticizes the cited work) are indiscriminately integrated in the citation counts. But, as several studies reported in (Bornmann and Daniel 2008) have shown, these kinds of citations are very rare (generally less than 10% of occurrences). This particular issue is addressed in section 6, based on our own data.

Recent work by Simone Teufel and her colleagues has addressed the automatic classification of citations, based on the citation context. Their approach relies on a large number of linguistic cues, ranging from specific keywords and patterns to the relative position of the citation in the article (see Teufel et al 2006 for more details). They achieve very good results on their corpus of computational linguistics articles, using a 12-category typology and a machine-learning method based on the manual annotation by several annotators. This specific work has inspired our own, as presented in section 4.

- Garfield 1962 (general)
  - ☐ Paying homage to pioneers.
  - ☐ Giving credit for related work (homage to peers).
  - ☐ Identifying methodology, equipment, etc.
  - ☐ Providing background reading.
  - ☐ Correcting one's own work.
  - ☐ Correcting the work of others.
  - ☐ Criticizing previous work.
  - ☐ Substantiating claims.
  - ☐ Alerting to forthcoming work.
  - ☐ Providing leads to poorly disseminated, poorly indexed, or uncited work.
  - ☐ Authenticating data and classes of fact (physical constants, etc.).
  - ☐ Identifying original publications in which an idea or concept was discussed.
  - ☐ Identifying original publication or other work describing an eponymic concept or term.
  - ☐ Disclaiming work or ideas of others (negative claims).
  - ☐ Disputing priority claims of others (negative homage)
- Krampen and Montada 2002 (psychology)
  - ☐ Direct reference to an empirical finding in the cited document
  - ☐ Simple mention (of the type "compare here also," "see also," "see, for example") without any further more specific reference to the cited document
  - ☐ Direct reference to a theory or concept in the cited document
  - ☐ Direct reference to a method in the cited document
  - ☐ Overview citation (of the type "for an overview, see here," "see summary in") without any further reference to the cited document
  - ☐ Use of a data collection method (such as a test) taken from the cited document
  - ☐ Word-for-word quotation of text in the cited document
  - ☐ Use of a statistical method taken from the cited document
  - ☐ Substantial, theoretical, or methodological critique of the cited document
  - ☐ Use of a table, figure, or list taken from the cited document
  - ☐ Other citation type (for unclear citations)
- Teufel et al. 2006 (computational linguistics)
  - ☐ Weakness of cited approach
  - ☐ Contrast/Comparison in Goals or Methods (neutral)
  - ☐ Author's work is stated to be superior to cited work
  - ☐ Contrast/Comparison in Results (neutral)
  - ☐ Contrast between 2 cited methods
  - ☐ Author uses cited work as basis or starting point
  - ☐ Author uses tools/algorithms/data/definitions
  - ☐ Author adapts or modifies tools/algorithms/data
  - ☐ This citation is positive about approach used or problem addressed (used to motivate work in current paper)
  - ☐ Author's work and cited work are similar
  - ☐ Author's work and cited work are compatible/provide support for each other
  - ☐ Neutral description of cited work, or not enough textual evidence for above categories, or unlisted/unknown citation function

Figure 1: Example of citation function categories

2.3 Citation content analysis

A third and last set of ongoing researches on citations also focuses on the qualitative characterization of individual citations, but is concerned with the extraction of content from citation contexts. The objective of this approach is to (generally automatically) extract relevant information about a reference through the analysis of what citing authors relate about it. The most straightforward application is the extraction of keywords that can be used for information retrieval in databases, but also to classify or summarize publications. This is an interesting

alternative to traditional keyword extraction methods, that are applied to the content of the article itself: in the case of citation content analysis, the relevant content is produced by citing authors. The interest of this method can be clearly seen in the following example contexts taken from (Ritchie et al 2006), where the target keywords (in boldface) are obviously relevant descriptors of the cited references (underlined):

- *Such estimation is simplified from **HITS algorithm** (Kleinberg, 1998).*
- *For a **comparison to other taggers**, the reader is referred to (Zavrel and Daelemans, 1999).*

The most common method for extracting such keywords from citation contexts is to look for frequent terms in a collection of contexts referring to the target cited work (see also Schneider 2004). Of course, this requires the availability of such data, and it can only be applied to references that are cited a large number of times, in order to successfully apply statistical measures to candidate keywords.

2.4 Overview of needed data and processing

All three approaches described above require access to specific data. The widest range of useful data comprises:

1. the identification of the analysed article's features (author(s), date, title, journal, etc.);
2. the list of references extracted from the article, and their specific features (author(s), date, title, journal);
3. the citation context(s) for each reference, extracted from the analysed article;
4. the matching of references across different articles.

These different aspects are illustrated in figure 2.

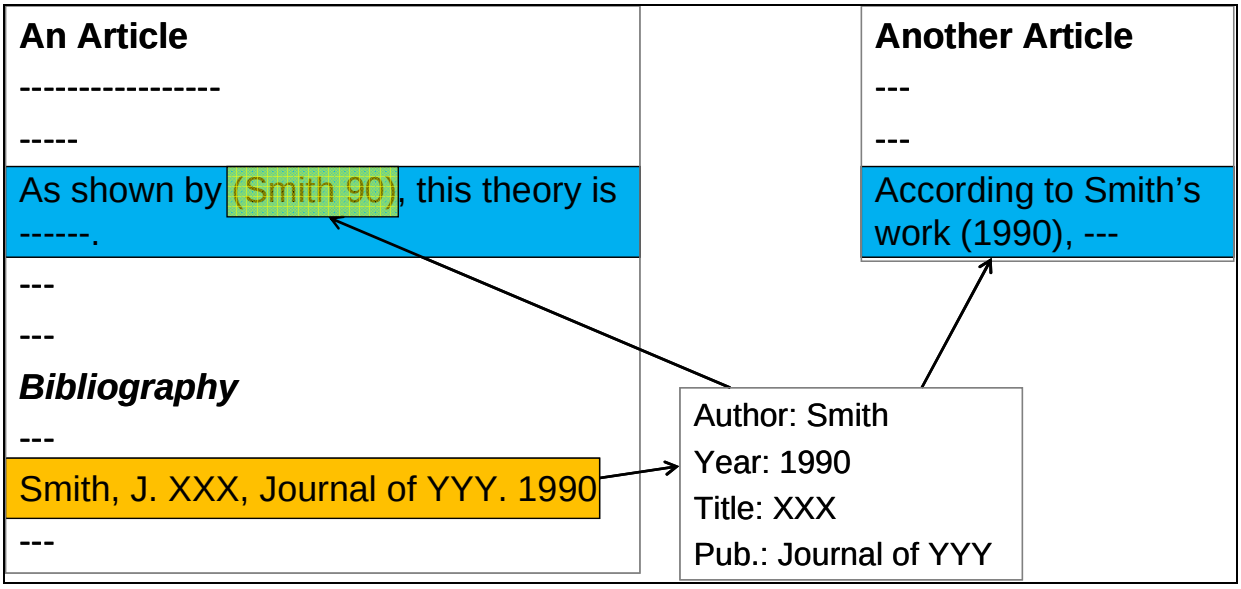

Figure 2: Overview of data used for citation analysis

Bibliometrics only needs items 1, 2 and 4. Citation classification needs 1, 2 and 3, while citation context analysis needs all of them.

Most citation studies use already parsed data. This is obviously the case for bibliometrics, where most of the efforts are in fact the building of the citation indexes themselves. Citation content analysis can be performed on the data provided by citation indexes but it relies most of the time on the availability of full-text articles, in order to have access to citation contexts. The fourth item in the above list is in fact the most difficult to obtain, as it requires a kind of normalisation of cited references, which can be extremely difficult across different journals with an important variation across bibliographic styles. Methods relying on the access to several different citation contexts for the same reference can of course only be applied to frequently cited work.
All these aspects show how difficult it is (in our case) to study the citations in French Humanities and Social Sciences. The main difficulty comes from the existing citation indexes, which have a very poor coverage of HSS fields, and nearly none of French language publications. Although the availability of online publications allows access to a large number of articles, all of the processing described above had to be done specifically for this project. The first consequence, given the difficulties, is that we could not have access to any connected data, i.e. that we had to work with individual citation contexts without any possibility to identify that two citations refer to the exact same reference.

3 Corpus and processing

As noted above, no data was made available through citation indexes that corresponded to our target (French Humanities and Social Sciences publications). However, the availability of such publications is not a problem anymore, thanks to the development, in recent years, of publication portals. Several websites provide (more or less freely) a large number of such publications, covering different disciplines and providing full text articles.
Below is the list of the main portals providing French language HSS publications:

- **Revues.org** (`www.revues.org`): This portal has been developed by the *Centre pour l'édition électronique ouverte* (CLEO, CNRS). It provides more than 180 different journals in free access, all belonging to the HSS field, and most of them written in French. Articles are available in XHTML format.
- **Cairn** (`www.cairn.info`): This portal originates from the joint efforts of French publishers (Belin, De Boeck, La Découverte and Erès). It now offers more than 150 HSS journals published by more than 40 different editors. Access to this database has to be paid for, and articles are in XHTML or PDF format.
- **Erudit** (`www.erudit.org`): This portal is run by a consortium of Canadian universities from Québec (Université de Montréal, Université Laval and Université du Québec à Montréal). The Erudit website provides access to several kinds of HSS and life sciences documents (journals, books, conference proceeding, theses). Documents are in XHTML and PDF format.
- **Persée** (`www.persee.fr`): This portal has been created by the French Ministry of Education and Research. Its main goal is to give access to older issues of HSS journals. Its content is also available through revues.org.
- **HAL-SHS** (`halshs.archives-ouvertes.fr`): This open archive, created by the CNRS, is dedicated to the publication of preprint material by their authors. Formats and sources vary widely, according to the authors' choice.

Considering the availability of data, their format, and various other reasons, we selected our first corpus from Revues.org, whose management is also part of the TGE-Adonis consortium. Another interesting aspect of this portal is that it follows the recommendations of the Open

Archive Initiative (OAI, *www.openarchives.org*), an international joint effort to normalise access to scientific publications. An important consequence of this initiative is the availability of a network protocol for querying the database (OAI Protocol for Metadata Harvesting), which can easily be automated in order to retrieve documents based on a number of descriptive features, such as language, date, disciplines, etc.

Our corpus has the following characteristics: it consists of 612 articles, published in 11 journals. It comprises a total of 8,105 References and 10,482 identified citations. It totals 4.5 million words. The disciplines covered are psychology, pedagogy/education, linguistics, ethnology, geography, sociology, history, law, anthropology, and philosophy. Several of the journals are interdisciplinary, which explains this diversity of fields.

In order to perform an automated analysis of this data, we applied the following processes to each article in our corpus:

1. Reference extraction: this has been done with ad hoc tools, based on the XHTML structure of the articles, where the "bibliography" section is specifically marked.
2. Reference parsing: we loosely adapted to our needs the ParaCite tools (which have been developed for English for the CiteSeer digital library (`citeseer.ist.psu.edu`). However, we limited the parsing to the extraction of the authors' names and the publication year.
3. Identification of citation: given the authors' names and publication year, we developed specific tools to mark up in the text the corresponding citations.
4. Parsing: we ran Synapse Développement's Cordial Parser (`www.synapse-fr.com`) on the articles' text. This parser provides state-of-the-art part-of-speech tagging, chunking and syntactic analysis.
5. Integration in the GATE platform: in order to query our data and mark up specific patterns, we used the GATE natural language processing platform (`gate.ac.uk`). This platform has the advantage to provide access to both the original structural mark-up from the article, in addition to the results of the linguistic analysis. It also provides specific tools to define patterns through the JAPE language.

XML format is used throughout all processing stages, both to store and give access to the structure of the articles, linguistic data and citation annotations. The following two sections present the specific experiments we designed using this environment.

4 First task: identifying "important" citations

4.1 Overview

Our initial objective was to follow Teufel and her colleagues' tracks and to apply natural language processing techniques to automatically identify the function of each citation in our corpus (Teufel et al. 2006). As presented in section 2.2, our first step was to adopt a citation function typology that could be applied to our data, i.e. to every discipline within HSS. Several attempts have been made to propose a generic typology, with a few top-level categories: negative/criticism, positive/praise, use of method/data/definition, neutral.

As said above, negative functions are extremely rare, and the few which we could identify manually were quite discouraging, making use of very indirect criticism, and sometimes irony, which we could not imagine to be able to catch (see section 6 for more details and examples on this topic). Citations referring to the use of data, methods or similar materials from the cited work were also very rare in our first experiments, as many HSS publications do not relate the use of tools or data whatsoever. Regarding the use of definitions from cited work, it is covered by our second objective, described in section 5.

The only satisfactory distinction we could identify was the one opposing “background” and “important” citations. This dichotomy can be seen in every citation function typology that has been proposed, although it is called by different names.
Here are a few examples of what we call “background citations”:

- Plusieurs études ont montré que les mères favorisent les phrases courtes (Brown et Bellugi, 1964 ; Drach, 1969 ; Lord, 1975 ; Moerk, 1975 ; Nelson, 1973 ; Newport, 1975 ; Phillips, 1973 ; Sachs, Brown et Salerno, 1972 ; Shatz et Gelman, 1973 ; Snow, 1972, 1977).
  ***Translation**: Several studies have shown that mothers prefer short sentences (Brown et Bellugi, 1964; Drach, 1969; Lord, 1975; Moerk, 1975; Nelson, 1973; Newport, 1975; Phillips, 1973; Sachs, Brown et Salerno, 1972; Shatz et Gelman, 1973; Snow, 1972, 1977).*

- Dans les sociétés occidentales, les auteurs (Caroll, 1974 ; Montandon et Crettaz, 1981 ; Lombardo, 1989 ; Chauvenet et al., 1994 ; Crouch, 1995 ; Crawley, 2004 ; Chauvenet, 2006) ont montré que ce groupe [...]
  ***Translation**: In western societies, the authors (Caroll, 1974; Montandon et Crettaz, 1981; Lombardo, 1989; Chauvenet et al., 1994; Crouch, 1995; Crawley, 2004; Chauvenet, 2006) have shown that this group […]*

- On sait maintenant (voir par exemple Perfetti, 1985) que la différence entre bons et mauvais lecteurs porte surtout [….]
  ***Translation**: We now know (see for example Perfetti, 1985) that the difference between good readers and bad ones lies mostly […]*

These citations are associated to the following cues:
- grouped in enumerations;
- in parentheses, in footnotes, i.e. less integrated in the sentence;
- located in the first parts of the article (introduction);
- in co-occurrence with specific marks: “the authors”, “see for example”, “see also”, etc.

We call “important” citations the other cases, of which here are a few examples:

- En cela, nous allons dans le sens des considérations de J.-P. Bronckart (1994) sur la double nature du genre.
  ***Translation**: In this, we follow J.-P. Bronckart (1994)’s considerations on the double nature of the genre.*

- C'est ainsi que nous avons établi, à la suite de Bogaards (1988), une distinction théorique entre ‘processus' et ‘stratégie’.
  ***Translation**: That’s how we established, following Bogaards (1988), a theoretical difference between ‘process’ and ‘strategy’.*

- C. Dévelotte (1989) se propose de mettre en évidence les stratégies individuelles de lecture mises en œuvre par des apprenants-lecteurs en F.L.E pour reconstruire le sens d'un article de presse.
  ***Translation** : C. Dévelotte (1989) proposes to identify the individual reading strategies by French language learners to rebuild the meaning of a news article.*

These citations are associated to the following features:
- isolated (i.e. not included in an enumeration);

- integrated in the sentence: as subject, object, possessive noun phrases, etc.;
- in co-occurrence with marks of the author's implication, such as first person pronouns.

Our hypothesis, following the study of these phenomenons, is that background or perfunctory citations are much less integrated than important ones in the author's speech. This integration feature is closely related to the observations made by Swales (1990), who makes a clear distinction between integral and non-integral citations, but without linking these aspects to any kind of importance or citation function.
This led us to propose a citation integration scale, in which we organise a set of different cues, as shown in figure 3 below.

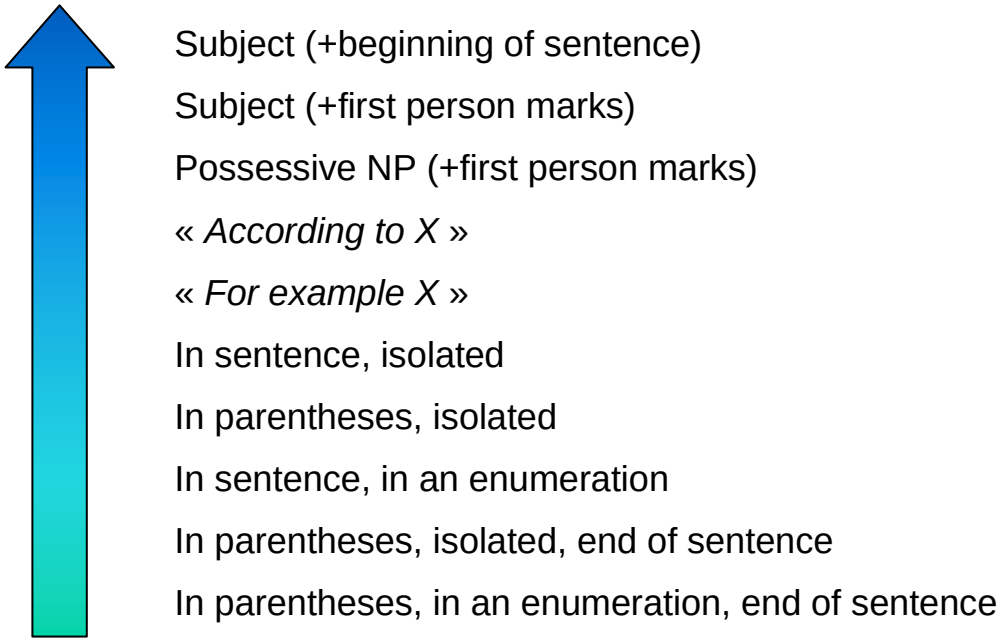

Figure 3: Proposition for a citation integration scale

In this scale, the upper configurations of contextual features are associated to important citations, while lower configurations are deemed associated to background citations.
Thus redefined, our task consists in the automatic extraction of the features mentioned above, and in the definition of decision rules to classify each citation in one of the two categories.

4.2 Automatic features extraction

Using the JAPE language provided by the GATE platform, we were able to design specific patterns that identified, for each citation in our corpus, a set of 8 attributes. These 8 characteristics are described below.

1. *Enumeration versus isolated citation*
   This feature is based on the number of citations surrounding the target. If one or more citations are in the immediate vicinity, it is considered to be part of an enumeration. Otherwise, it is considered to be isolated. The overall ratio is 25% in an enumeration against 75% isolated (for this features and the following, we could identify significant variations across journals and disciplines, but have not yet investigated their scope and meaning).

2. *In the sentence versus in parentheses*
   This second feature is also very simple to compute, as it consists in the presence/absence of parentheses in the target citation's surroundings. Around 20% of citations are integrated in the sentence, versus 80% in parentheses.

3. *Location in the sentence*
   Independently from the previous feature, we computed the general location of the target citation in the sentence: beginning, middle or end, based on the structural mark-up and the automatic sentence tokenization performed by the parser. The beginning and end of a sentence are relative (and not absolute) locations, which means that the method allows a few words to be present before or after the target citation. Overall, 8% of citations are found at the beginning of a sentence, and 50% at the end.

4. *"According to <cit>"*
   We specifically searched for patterns that are equivalent to the English general "*according to*" scheme. More precisely, we selected the following prepositions: *selon*, *pour*, *chez* and *d'après*. This feature is one of the least productive in our study, as it co-occurs with about 1.5% of citations.

5. *First person marks*
   This features consists in searching, in the sentence containing the target citations, for explicit marks of the first person (i.e. the article's author), more specifically pronouns and possessive determiners (*je*, *me*, *moi*, *nous*, *mon*, *mes*, *nos*, *notre*, etc.). Overall, around 3.5% of citations appear in such contexts.

6. *Syntactic function*
   For citations integrated in the sentence, we used the parser's results to check whether the citation was the subject of the main verb (other syntactic functions, such as direct object, were not frequent enough to be taken into account). Around 5.5% of all citations are used as subjects.

7. *Exemplification*
   This feature is related to specific contexts equivalent to the English "For example" and "See also", that confer a status of example to the target citation. Around 1.6% of citations appear in such contexts.

8. *Possessive noun phrases*
   This feature focuses on noun phrases in which the target citation has a possessive function, such as in "*<cit>'s work*". We built a list of 37 nouns and noun phrases that refer to the cited work, for example *travaux*, *article*, *approche*, *thèse*, *théorie*, *modèle*, etc. Thus defined, this feature has been matched with 1.6% of all citations.

Each feature was the target of a specific manual evaluation on a sample of our corpus, and the results were very satisfactory, with recall and precision scores both of 95%, which means that our automatic features are reliable.

Once all these features have been defined and calculated, we are able to design and test simple decision rules in order to propose an automatic classification of every citation in our corpus. Due to the low frequency of some of the cues, our first automated decision strategies will be

very straightforward and will only focus on non-marginal features, i.e. in sentence versus in parentheses and isolated versus in an enumeration. Naturally, more complex scenarios can be defined, and give way to a continuum of importance, but the mandatory next step was to confront our hypotheses to human decisions.

4.3 First experiment and preliminary results

In order to test the method described above, we asked different annotators to assess the importance of the citations in a set of articles. Due to logistical difficulties independent from our will, we could only achieve partial results, and had to limit our study to 2 different articles, 3 different annotators, and only one disciplinary field. More precisely, we chose the field of language didactics, and asked 3 Master students to assess a total of 107 citations corresponding to 69 references.

The 2 articles were selected to be different, with article 1 being a general, state of the art presentation of different approaches to a research topic, and article 2 being a more focused and specific study. The 3 annotators were also chosen for their different levels of knowledge of the field. Annotator 1 is writing her master's thesis in this specific area, and can thus be considered a specialist. Annotator 2's area of expertise is related, but not too closely, to the articles' topic, while Annotator 3, while still a linguistics student, is specialised in another part of the field. All three annotators were therefore able to understand the two articles, but had very different knowledge of the cited works.

The task was described to the annotators in the following terms: "For each reference cited in this article, do you find it to be important or related to background knowledge?". In order to help the annotators in their decision, they were given a list of Garfield's citation functions as example of each category.

Feedback from the annotators showed that the task was found to be difficult, but all three were nevertheless satisfied with their ability to discriminate citations with a number of understandable hesitations for a few items.

We computed the inter-annotator agreement scores (i.e. Cohen's kappa) for each pair of annotators, as shown in table 1.

Table 1: Inter-annotator agreement scores for the important/background categorisation of citations

Kappa	**Annotator 1 vs 2**	**Annotator 1 vs 3**	**Annotator 2 vs 3**
Article 1	0.84	- 0.12	- 0.05
Article 2	0.59	0.20	0.26

The kappa score is a well-known measure of inter-annotator agreement based on the ratio of tested items on which two annotators agree, while taking into account the expected results if the annotators answer at random. Its values are comprised between -1 (perfect disagreement) and 1 (perfect agreement), with 0 meaning an absence of agreement (i.e. the annotators achieve the same level of agreement as a random answering). There has been a long-time controversy about the thresholds that can be used to interpret the obtained Kappa values. Therefore, we will carefully limit our interpretation to a comparison between the scores obtained for different configurations.

Firstly, it appears that annotators 1 and 2 reach a high level of agreement (close to perfection for Article 1), while annotator 3 clearly takes different decisions for an important number of citations. This is a clear (but not surprising) clue that knowledge of the disciplinary field is an important factor for this task (annotator 3 having no prior knowledge of any of the tested references).

Secondly, the nature of the research article is also a factor for this decision task, with the second article leading to better agreement with non-specialist. Article 2 follows a more classical scheme, i.e. it describes an experiment (hypothesis, data, methods, results, etc.), while Article 1 covers a wide area of scholarly knowledge, with seemingly less distinctions between background citations and "important" ones. In the case of Article 1, the identification of importance relies more on the readers' knowledge.

We then confronted the human annotations with the results from our automatic classification. Our first approach was to declare a citation as "important" whenever it appears in the sentence, and isolated. This means that every citation in parentheses or in an enumeration is considered as "background". In case of several different citations for the same reference, the "important" decision prevails. The accuracy scores shown in table 2 are the percentage numbers of references for which both the annotator and the automatic decision system agree.

Table 2: accuracy scores of automatic classification, first rule

Accuracy (%)	**Annotator 1**	**Annotator 2**	**Annotator 3**
Article 1	38	31	61
Article 2	63	63	74

Once again, the resulting measures vary widely from one configuration to another. Accuracy reaches much higher scores for the second (more standard) article than for the first one and for the third annotator (novice) compared to the first two.
The most important result at this stage is that the automated classification measure is seemingly close to the default reader's behaviour when no specific knowledge of the field interferes with the task. Obviously, this method is less successful for more expert readers.

A closer look at the cases of disagreement (mostly for annotators 1 and 2) tells us that our decision rule is too restrictive, and that these annotators identify more important references than the automatic decision system. According to this, we designed another decision rule, which this time considers any isolated citation as "important", whether it appears in parentheses or in the sentence. The results are shown in table 3.

Table 3: accuracy scores of automatic classification, second rule

Accuracy (%)	**Annotator 1**	**Annotator 2**	**Annotator 3**
Article 1	38	31	31
Article 2	78	70	52

This more tolerant rule reaches satisfactory accuracy scores for the first two annotators on the second article, with a lower value for the novice annotator. Scores remain very low for the first article for all three annotators. This method seems to capture more precisely the behaviour of expert readers on a standard article.

These results, although globally inconclusive due to the size and variety of the test data, lead us to the following (temporary) remarks. Firstly, the task of assessing the importance of citations is a difficult one. This was already found to be so by previous studies (Hanney et al 2005 obtained similarly low inter-annotator kappa scores). It can however reach satisfactory agreement levels in some good conditions, i.e. with expert readers and more standard publications, as was the case for Teufel et al 2006 (kappa scores of more than 0.7 were reached on an even more complex task).

Secondly, our proposed citation integration scale needs to be further refined and tested, as many of our cues have yet to be taken into account. If it appears quite difficult to obtain an overall satisfactory result using this approach, it still has shown encouraging results when applied in a favourable configuration, *i.e.* when compared to a non-expert reader or a standard-structured research article.

5 Second task: extracting citation keywords and concepts

The second objective assigned to this exploration of citations in French HSS articles is closely related to the third mentioned approach of citation analysis, in which the content of citation contexts is specifically investigated. As described in section 2.3, most similar approaches use term extraction techniques to extract relevant keywords from citation contexts, and generally do so with the help of frequency filtering across different contexts concerning a target reference. Due to our lack of cross-referenced data, we had to follow a different approach.
In HSS articles, a significant number of citations seem to have a specific function, which can be expressed as the borrowing of a term or concept from the cited work. The following examples have been extracted from our corpus:

- […] un modèle langagier représentant approximativement un des stades du « **processus de complexification** » (Corder, 1978) d'une langue institutionnalisée […]
  ***Translation**: […] a language model which approximately represents one of the stages of the « **complexification process** » (Corder, 1978) of an institutionalised language […]*

- […] on interprète souvent ce passage de la **« revanche à la contrainte »** (Cohen, 1985, 76) en y voyant un développement profondément positif.
  ***Translation**: […] this evolution **"from revenge to constraint"** (Cohen, 1985, 76) is often interpreted as a deeply positive development.*

- […] ce que Yau (1992 :146) appelle le **paradigme de coordination**.
  ***Translation**: […] which Yau (1992: 146) calls **the coordination paradigm**.*

- L'individualisation des trois formations discursives de la déviance criminalisée que je propose se limitera, pour l'essentiel, à ce que Foucault nomme **la différenciation primaire des objets** (97).
  ***Translation**: The individuation of the three discourse formations of criminalised deviance that I propose will be limited, mostly, to what Foucault names **the primary differentiation of objects** (97).*

In each of these examples, the citation's main function appears to be the introduction of a concept which is borrowed from the cited reference, or to use a new term coined by the cited author in the citing author's own speech. This function has already been identified as such in the different typologies of citation functions, most of the time regrouped with similar ones, such as Garfield's "*Identifying original publication or other work describing an eponymic concept or term*", Krampen and Montada's "*Direct reference to a theory or concept in the cited document*" or Teufel's "*Author uses tools/algorithms/data/definitions*". In HSS publications, these specific uses of citations seem to appear quite frequently, concepts and terms being much more borrowed in more theory-oriented articles than data or methods.
From a linguistic point of view, the cues of such contexts are quite easy to identify, and fall into the two following configurations:

1. Quotation marks or typeface (italics) to highlight the term/concept;
2. Definition-like patterns, with the use of naming verbs (*appeller*, *nommer*, *dénommer*,

utiliser/proposer/employer le terme/mot/expression, etc.).

The first one has of course to be distinguished from longer quotations, where a whole passage of the cited reference is reproduced. Our target patterns here are limited to noun phrases or very short sequences of text, and not whole sentences or paragraphs.
The second one can be seen as a more specific aspect of definitional contexts, as they have been thoroughly described and formalised by (Rebeyrolle and Tanguy 2002). In our case, we had to focus on naming definitions involving a citation or author.
Given these first observations, we were able to develop specific JAPE grammars to identify these specific contexts, and to extract the borrowed terms, that can be associated to the corresponding reference. These local grammars rely on the syntactic parsing for NP identification, and make use of several word lists. Overall, we managed to extract 808 such terms, with an estimated precision score of 80%.
A number of errors are due to the patterns themselves that encounter some difficulties when dealing with some specific contexts. Apart from a few parser errors, these difficulties are related to:

- the polysemy of some verbs such as *nommer* (meaning both *to name* and *to appoint*) and *appeler* (*to call*);
- the inherent polysemy of quotation marks, which some authors use as modals, such as in:
  - La libération conditionnelle devient avant tout une mesure de gestion peu coûteuse de condamnés objectivés comme des « déchets » […] (Simon, 1993, 142 et 259; Simon et Feeley, 2003, 99; 1994, 193 ; Lynch, 1998).
    ***Translation**: The release on parole becomes mostly a low-cost solution to manage convicts considered like "waste" […] (Simon, 1993, 142 et 259; Simon et Feeley, 2003, 99; 1994, 193; Lynch, 1998).*
- the difficulty of some contexts, such as this one, where two different cited terms appear, the second being a translation or reformulation of the first one:
  - Le **« hourrah football »** (Sansot, 1991, p. 142), **le football des trottoirs** (Therme, 1995), investit la rue […]
    ***Translation**: The "**hourrah football**" (Sansort, 1991, p. 142), **the sidewalk football** (Therme, 1995) moves into the street […]*
    In this case, the first term (*hourrah football*) is successfully extracted, but not the second one (*football des trottoirs*), as it does not appear inside quotation marks nor with a specific cue phrase.

Interestingly enough, even in our small corpus, this term extraction method allowed us to find very similar contexts associating a definition to the same cited author:

- Certains auteurs comme Lazarus et Launier (1978) définissent le ***stress*** comme : "un processus qui apparaît quand les exigences environnementales dépassent les capacités de réponse de l'organisme".
- Certains auteurs, Lazarus (1966) définissent le ***stress*** comme : « un processus qui apparaît quand les exigences environnementales dépassent les capacités de réponses de l'organisme. »
  ***Translation (for both examples):** Some authors, like Lazarus and Launier (1978) define stress as "a process that occurs when environmental requirements exceed the organism's response capacity".*

These two excerpts come from entirely different articles, with different authors and in two different journals. Indeed, it appears that this definition of "stress" is a very commonly cited one, across different authors and disciplines, as confirmed by querying a Web search engine, leading to other very similar excerpts, such as this one:

- LAZARUS (1966), LAZARUS et LAUNIER (1978) vont plus loin en définissant le stress comme un processus qui apparaît quand les exigences environnementales dépassent les capacités de réponse de l'organisme.
  Translation: *Lazarus (1966), Lazarus and Launier (1978) go even further by defining stress as a process that occurs when environmental requirements exceed the organism's response capacity.*

Although our goal is not, contrarily to other similar approaches, to automatically index or summarize the cited references, we find that such borrowed terms can be of interest to the user of a publications portal, and as such can be proposed to him as a reading aid. Further details are presented in the conclusion of this article.

6 Other aspects of citations

In this last section we address a number of issues regarding our experience of citation analysis, and discuss a number of possible developments.

6.1 Positive and negative citations

Many discussions regarding qualitative citation analysis concern positive versus negative citations. More precisely, the "negative" citation has often been raised as an argument against purely quantitative citation analysis as done by bibliometricians. Most studies of citation functions propose a category that covers a kind of criticism of the cited work. Although such criticisms can easily be found and understood in science, where objective evaluation procedures can be used to compare results or techniques (and therefore lead to the conclusion that the cited work is inferior, or sometimes simply wrong), things are more complicated in Humanities and Social Sciences.
While manually analysing our corpus, we kept an eye open to such negatively evaluated citations, but could only find a few, some of which are shown below:

- Par exemple, l'éminent criminologue Chaiken (2000, 1), s'inspirant d'une publication de Statistique Canada, indique que la criminalité dans ce pays est stable (ou bien Chaiken est de mauvaise foi, ou bien il a examiné trop rapidement les données).
  Translation: *For example, the eminent criminologist Chaiken (2000, 1), citing a publication by Canadian Statistics, indicates that the crime rate in this country is stable (either Chaiken is insincere, or he examined the data too quickly).*

- […] hypothèse formulée par Lenton (1989) et correctement rejetée par Hagan (1991)
  Translation*: […] hypothesis formulated by Lenton (1989) and correctly rejected by Hagan (1991).*

It appears clearly that the identification of such negative citations is extremely difficult. In the first example, irony is used to express the negativity (*eminent criminologist*), and the very straightforward accusation (in the following parentheses) is quite unexpected in such a context. In the second example, it is the conjunction of a positive adverb (*correctly*) with a negative verb (*rejected*) that attributes a negative interpretation to the first citation. Both phenomenons are extremely complex, and current developments in automated opinion analysis cannot reach so high a level of subtlety.

On the other hand, positive evaluation of citations is most of the time directly expressed, as in the following example, were a positive evaluative adverb (*bien*) is explicitly associated to the citation.

- […] le mouvement de communautarisation de la justice qui, comme l'a bien analysé Crawford (2001 ; 1997), consiste à […]
  ***Translation**: […] the communautarisation of justice which, as Crawford (2001; 1997) correctly analysed, consists in […]*

However, with only one side of the spectrum accessible by automatic methods, the distinction between positive and negative citations is not in our current list of objectives.

6.2 From citation profiles to articles categorisation

As shown in our experiment with human annotation, it appears that citation analysis is highly dependent on several factors, among which the kind of article in which these citations are found. Our comparison between a state-of-the-art review and a more standard research report incites us to investigate the possible (automatic) means to classify such articles. An interesting lead is to examine the *citation profile* of a research article, which can be done in a number of ways. The number of references, the frequency of the corresponding citations, and their locations in the article can all be seen as clues to the general status of the article. The diagrams in figure 4 show the citation profiles of 4 different articles, based on the locations of citations in the text. In this first approximation, articles have been divided in 5 parts of equal length, and the number of citations in each part has been calculated. The diagrams show different profiles, among which the upper-left corresponds to a standard research report with a decreasing density of citations, most of which are located in the first parts to introduce the subject, hypotheses, etc., the latter parts of the article concerning the presentation of original work and no citations at all. The flatter lower-left profile, with more evenly distributed citations is a state-of-the-art review. The two profiles on the right side are extracted from more unusual articles, with very few citations (1 or 2), that can be located on either of the article's ends (upper right) or only in the middle (lower right).
We are currently investigating the possible exploitation of such profiles, but this requires further development, as we intend to use time-series analysis and data-mining techniques to identify different clusters of articles on the basis of this data. Our long-term objective is to implement an automatic classification technique to identify the generic category of an article. Once this is done, it will be possible to design and apply different strategies in order to resolve other tasks, such as the categorisation of individual citations described in section 4.

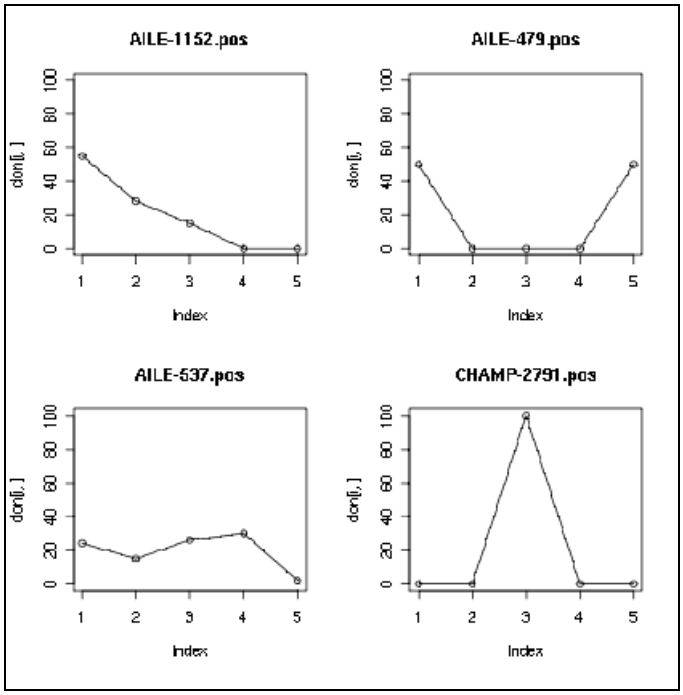

Figure 4: Example citation profiles

7 Conclusion

Citation analysis can be performed in a number of ways, to achieve different objectives: evaluating research, investigating the structure of an academic field, designing specific information retrieval tools, exploring the linguistic characteristics of research articles, etc. All these approaches need a large amount of data and extensive processing. Most of all, they rely on a dense connection of academic publications through citation links. Dealing with French Humanities and Social Sciences, we encounter a number of difficulties, ranging from the absence of French data in citation indexes to the lack of previous studies of these disciplinary fields. Nevertheless, we managed to propose two methods: the first one is an automated classification of citations according to their "important" versus "background" status, and the second one concerns the extraction of terms and concepts that have been borrowed from the cited work. Although the preliminary experiments have shown that the techniques still need some refinement, we believe that the results can be of use to a reader, by providing additional information on a research article. For example, the processing methods described in this paper allow us to automatically build a synthetic annotated bibliography, as shown in figure 5. In this excerpt, the original list of references has been filtered, and only the "important" references are shown, as decided by the method described in section 4. The citation contexts (sentences) are shown in the right-side column, where the relevant features are underlined. In addition, the associated terms, as extracted by the method described in section 5, are shown in boldface, as it is the case for the third reference ("*les procédures*"). This kind of synthetic table can be seen as a reading aid, and its deployment on a publication portal will be examined in the follow-up of this project.

TITLE: Quelques sources de variation chez les enfants

AUTHOR: Jisa, Harriet et Richaud, Frédérique

REF: AILE, vol 4, 1994

Reference	Context and relevant features
Ref 7: BERNSTEIN, B. (1971). Class, Codes and Control, Volume 1. London : Routledge and Kegan Paul.	A propos de l'utilisation différentielle de noms et de pronoms, il nous semble important d'examiner le travail de **Bernstein (1971)**.
Ref 10: BLOOM, L., P. LIGHTBROWN & L. HOOD (1974). « Imitation in language development: if, when and why ? », Cognitive Psychology 6 : 80-420.	**Bloom, Lightbrown and Hood (1974/1991)** ont examiné l'utilisation de l'imitation dans le discours spontané de six enfants.
Ref 18: BRUNER, J. (1983). Le développement de l'enfant : Savoir faire, savoir dire. Paris: Presses Universitaires de France.	Il se peut toutefois que l'étude des corrélations, qui porte sur ce que **Bruner (1983)** appelle **les procédures**, c'est-à-dire les adaptations faites par le partenaire plus compétent, masque la véritable importance du travail sur le code dans la co-participation.

[…]

Figure 5: Example annotated bibliography

Acknowledgements

The authors wish to thank the colleagues who are or have been involved in the RHECITAS project (in alphabetical order):
Farah Benamara (IRIT), Dominique Besagni (INIST), Cécile Fabre (CLLE), Lydia-Mai Ho-Dac (CLLE), Josiane Mothe (IRIT), Sophie Nègre (Synapse), Marie-Paule Péry-Woodley (CLLE), Clément Picou (Synapse) , Marjorie Raufast (CLLE), Josette Rebeyrolle (CLLE).

References

- Bornmann, L. and H.-D. Daniel (2008). What do citation counts measure? A review of studies on citing behavior, *Journal of Documentation*, 64:1, 45-80.
- Garfield, E. (1962). Can citation indexing be automated? *Essay of an Information Scientist, 1*, 84-90.
- Hanney, S., I. Frame, J. Grant, M. Buxton, T. Young, and G. Lewison (2005). Using categorisations of citations when assessing the outcomes from health research. *Scientometrics , 3* (65), 357-379.
- Krampen, G. and L. Montada (2002). *Wissenschaftsforschung in der Psychologie*, Hogrefe, Götingen.
- Moed, H. F. (2005). *Citation analysis in research evaluation*, Springer.
- Rebeyrolle, J. and L. Tanguy (2002) *Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires.* Cahiers de Grammaire, 25.

- Ritchie, A., S. Teufel and S. Robertson (2006) *How to find better index terms through citations*. In: Proceedings of the Workshop "Can Computational Linguistics Improve Information Retrieval?", at ACL/COLING-2006, Sydney, Australia.
- Ritchie, A., S. Teufel and S. Robertson. (2008) *Using Terms from Citations for IR: Some First Results*, Proceeding of ECIR, 211-221.
- Schneider, J. W. (2004) *Verification of bibliometric methods' applicability for thesaurus construction*, PhD Thesis, Royal School of Library and Information Science, Aalborg.
- Swales, J. (1990). *English in academic and research settings*, Cambridge University Press.
- Teufel, S., A. Siddharthan, and D. Tidhar (2006). *Automatic classification of citation function*, Proceedings of EMNLP 6, 103-110.