informationresearch

VOL. 17 NO. 4, DECEMBER, 2012

A method for the analysis of information use in source-based writing

Eero Sormunen
School of Information Sciences, FIN-33014 University of Tampere, Finland
Jannica Heinström
Research Collegium, FIN-33014 University of Tampere, Finland
Leena Romu and Risto Turunen
School of Information Sciences, FIN-33014 University of Tampere, Finland

Abstract
Introduction. Past research on source-based writing assignments has hesitated to scrutinize how students actually use information afforded by sources. This paper introduces a method for the analysis of text transformations from sources to texts composed. The method is aimed to serve scholars in building a more detailed understanding of how students work with sources, for example, in paraphrasing, summarising and synthesising information.
Method. The proposed method is introduced by presenting its domain, procedure and justifications, and by sketching a coding framework for text transformations. The characteristics of the method are demonstrated by reporting a case study: the use of information in seventeen Wikipedia/wiki articles written in a collaborative assignment by upper secondary school students.
Analysis. The domain of the method is represented by characterising its goal and application area. The procedure of the method is represented as an ordered set of operations and its use is demonstrated in the case study. The justifications of the method are addressed by discussing appropriateness, validity, reliability and efficiency issues related to the method.
Results. The findings of the case study demonstrate that new research questions can be answered by applying the method. In terms of research economy, the method is reasonably efficient. No major problems related to the validity and reliability of the method were observed.
Conclusions. The proposed method is a novel research instrument for the study of information use. It opens up interesting possibilities to analyse text transformations in source-based writing and expand our understanding of the core processes of information use.

Introduction

The lack of research, theoretical understanding and methodological development in studies on information use is often noted in library and information science (see e.g. Kari 2007; Savolainen

2009a and b; Vakkari 1997; 2008). One of the challenges of use studies is the concept of *use* and its definition (Kari 2007, 2010; Savolainen 2009a). Closely following are the methodological challenges of operationalising this concept.

The goal of our study was to enhance the tools available for the study of information use and knowledge construction by developing a method for the analysis of text transformations from sources to written texts. This type of text analysis has previously been used in reading-to-write studies (see Spivey 1997; Wiley and Voss 1999) but not in information studies. We prepared our approach in a pilot study the findings of which are reported elsewhere (Sormunen and Lehtiö 2011). In this paper we introduce an extended and revised method and the findings of a larger case study where the method was applied.

We argue that source-based writing is one of the most typical modes of information use and that it is crucial to understand how information derived from sources is processed in composing new texts. For example, in information literacy instruction it is not enough to warn students against copy-pasting. Rather we should illustrate in concrete terms how valid arguments are built in the creative scrutiny of sources and in the synthesis of information. Students also need to understand the contextual requirements of the task, such as the genre in which they write.

The paper is organized as follows: First, a short review on related research is given. Second, we introduce the proposed method by describing the domain (application area) for which it is designed, the procedure for using it and a coding scheme used in demonstrating the method. Then a case study exemplifying the method as a researcher's tool is reported. On the basis of the case study, we present justifications for the method by discussing its appropriateness, validity, reliability and efficiency.

Related research

In the literature review, we first look at information use in the context of learning. We then visit common definitions and research methods applied in studies of information use. Thirdly, experimental research done in reading-to-write studies is introduced. Finally, we present the summary of our pilot study

Studies on information use in a learning context

Previous research suggests a link between student motivation, depth of knowledge construction, and information use in classroom assignments. The analysis of written texts as expressions of topical understanding sheds light on both information use and knowledge construction as a process. Descriptive texts have been found to imply a more superficial knowledge construction process, while more analytical texts often reveal more sophisticated in-depth learning (Todd 2006).

Limberg (1999) observed that poorly performing students had a tendency to *fact finding* approaches which easily led to copy-pasting in independent learning tasks. These students are not interested in genuine inquiry but rather in collecting *right answers* from sources and transferring them to a research paper (Alexandersson and Limberg 2003). Similarly students with a surface approach to studying tend to gather information merely for task completion as opposed to striving for genuine learning (Heinström 2003; 2006). McGregor and Streitenberger (2004) observed that the levels of

copying and plagiarism were higher among students who concentrated on the format of the end product (*looking good*) than on the process of gathering and synthesising information for its content.

In the internet age, it is an obvious risk that students transfer information mechanically from sources to their own texts instead of transforming it in the cognitive process of knowledge construction. The least engaged students thereby fail to achieve learning goals in topical contents as well as in information literacies. Surprisingly, copy-pasting has rarely been studied by comparing sources used and texts written, except in the context of plagiarism (see McGregor and Streitenberger 2004; McGregor and Williamson 2005).

The concept of information use and methods for studying it

So far, the information science community has not agreed upon the definition of information use. Different conceptions of the term include interpreting *information use* as: information practice, information search, information processing, knowledge construction, information production, applying information and effects of information (Kari 2010). Information use is also interpreted in diverse ways by users themselves (Kirk 2002; Maybee 2006).

Kari (2007) suggests that the term information *outcome* might be a more appropriate term, as it includes both information use (what a person does with information) and information effect (the influence of information on a person). Most studies on information use within information science have focused on what Kari (2007) calls information effect. One example is the cognitive constructivism approach, which focuses on the process of knowledge construction (Savolainen 2009a). An alternative viewpoint is to regard information use as '*a process that is contextualized in action or practice*' (Savolainen 2009a). This notion emphasizes what people do with information rather than what information does to people (Kari 2007) and, in particular, underlines the importance of context (Savolainen 2009a).

Different conceptions and levels of information use calls for a variety of methodological approaches. The cognitive constructivism view is underlined, for example, in Brookes's fundamental equation (1980). Wilson (2000: 50) defines information use as '*the physical and mental acts involved in incorporating the information found into the person's existing knowledge base*'. Within this tradition the most common research methods aim to capture the influence of information by exploring expressions of (changing) knowledge through interviews or text analysis.

Through interviews Cole (1997) found that information use proceeds as a process from the first information encounter, which needs to be powerful enough to initiate a consequent information process, to the final effect of new insights. Studies on information use from a process viewpoint have also used a combination of text analysis and interviews (see e.g. Todd 1999), as detailed processing of information may be difficult to recall in an interview setting. A possible exception is particularly important insights which in fact may be recalled very specifically. For example, Cole's (1997) interview studies revealed that doctoral students could remember where on a page an influential information piece had been found.

Information use as a knowledge construction process often requires a longitudinal approach in order to capture changing knowledge structures. Todd's (1999; 2006) work illustrates this process in his studies on how consulting information changes knowledge structures. Todd analysed the understanding of a topic in a stepwise process by comparing evolving texts in a learning process over time.

In studies on plagiarism the focus in information use is on what people do with information (Kari 2007). Researchers of plagiarism in school assignments have mainly applied ethnographic methods to shed light on the phenomenon and extended the view by quantitative analysis of plagiarism and copying in students' research papers. McGregor and Streitenberger (2004) developed a five-level classification for the comparison of texts in student reports and sources used. This was also applied in a later study by McGregor and Williamson (2005). The authors contribute by introducing a model by which to compare the texts of written reports and used sources. The approach, however, only includes the dimensions of copying and plagiarism and ignores other aspects of information use in source-based writing.

Perhaps the best way to approach the daunting task of measuring such a wide and conceptually challenging concept as information use is to operationalize it in concrete terms within a specific context. Savolainen (2009a) underlines the importance of considering the context of information use. Expressions of information may vary depending on the discourse of a particular social setting and the pragmatic social purpose for which it is produced (Tuominen and Savolainen 1997). In our case study, the students' goal was to produce a factual encyclopaedic text. This meant that the students had to adhere to the rules of the genre.

Source-based writing as a reading-to-write task

Spivey (1997: 136) defines the reading-to-write task as a process whereby a person is concurrently in two roles: in the role of reader *building meaning* from a text and in the role of writer *building meaning* for a text. Kiili, Laurinen, Marttunen and Leu (in press) call the former process *meaning construction* and the latter, knowledge construction. The writer reads others' texts (sources) but also his/her own when composing it. Writing starts as a cognitive process while reading in form of planning how the sources can be used in the text to be written (Spivey 1997: 144-145). Rouet (2006: 91-92) raises an important challenge for building meaning from sources. Within the task constraints, the reader seldom has a chance to carefully read all documents to build complete meaning from them. Rather, he or she has to search for relevant pieces of information from documents and at the same time derive justified arguments from sources.

Making a synthesis across sources is more demanding than writing a summary of a single text (Davis-Lenski and Johns 1997, Mateos and Solé 2009). In *summarizing* a single text it is possible to maintain the structure of the original text. The *synthesis* of multiple texts requires an integrating idea of how to transform information from differently structured, even contradictory, texts into a new structure. The synthesis requires knowledge transformation to a greater extent than does making a summary (Mateos and Sole 2009; Segev-Miller 2004).

In the experimental settings used in reading-to-write studies (see Spivey 1997; Segev-Miller 2004; Wiley and Voss 1999), students are typically given two or more source texts and asked to write their own texts on the basis of their readings. In recent studies focused on new literacies, students have been given a similar task but unlimited access to internet sources (Kiili *et al.* 2008; Kiili, *et al.* in press). Researchers collect and analyse
data on the process and the resulting texts.

Spivey (1997: 149-163) gives a description of the classic method of reading-to-write studies which analyses the relationships between sources and texts written. She parsed source texts and texts written by students' into propositions called content units. On the basis of this semantic representation, for example, unique and overlapping contents of source texts as well as source texts and written texts could be identified. The analysis of texts at the level of propositions requires huge

resources as the number of texts increases and is not feasible in a typical school assignment situation of multiple writing topics and information sources.

Wiley and Voss (1999) introduce a more realistic approach in terms of research economics. They conducted a controlled experiment where students wrote essays on historical themes based on a given set of textbook and web-like sources. They classified the origin of each sentence in essays using a three-category scheme. A sentence was coded as,

1. *borrowed*, if it was taken directly or paraphrased from the sources
2. *transformed*, if it contained source-based and novel information combined or connected two or more pieces of information from sources
3. *added*, if it contained only novel information beyond the sources.

The scheme is quite simple but works as a point of departure in developing a more elaborated framework for text transformations.

Pilot study

A pilot study by Sormunen and Lehtiö (2011) introduced a method for the analysis of text transformations which could be applied in a setting where students are free to use any number of sources in writing their texts. The fact that students may not cite all used sources was taken into account and a plagiarism check was included in the method's procedure. A one-dimensional categorization for sentence-level text transformations was introduced: *copy-pasting; near copy-pasting; paraphrasing; summarizing from a single source and synthesising across sources.*

The use of the proposed method was demonstrated in a data set of eleven short Wikipedia articles and seventy sources used in writing them. One limitation of the coding scheme was that it merged two variables – the degree of paraphrasing and the degree of information synthesis – into a single dimension. The original coding scheme introduced no credibility measure for the use of sources. Further, the analysis was made totally on sentences excluding broader text structures such as paragraphs. A text paragraph could be a relevant unit of analysis, for example, in investigating the writer's overall tendency to synthesise information across sources.

Method proposed and text transformation dimensions

In this section, the domain of the proposed method and its procedure are described along the lines suggested by Newell (1969) for representing operational methods. The interim coding scheme used to demonstrate the method is introduced at the end of the section. The third main element suggested by Newell for representing methods – justifications - is presented in the discussion section after the case study.

Domain of the method

The aim of the proposed method is to reveal the extent to which students paraphrase, acknowledge sources and synthesise information in source-based writing – a typical assignment in information literacy instruction. The method is intended for scholars for the purpose of analysing source-based texts.

Procedure of the method

The procedure of the method proposed is illustrated in Figure 1. The first step is to collect all sources students used in writing their texts. In addition to cited sources, it is important to identify intentionally plagiarised or otherwise unacknowledged sources. Collecting materials for the analysis consists of five main steps:

1. All articles and sources cited in them are collected in electronic form if possible.
2. The articles are split into sentences and stored in an appropriate tabular format (e.g., Microsoft Word table).
3. The analyst poses a question for each sentence: 'From which sources and from which parts of those sources was this piece of information derived?' Matching is first attempted to sources appearing closest as in-text citations. In case of no match, searching is spread to all sources mentioned in the list of references. Search strategies are source type dependent. In electronic sources keyword searching is an efficient way to find the relevant fragments of text. If paraphrasing is extensive, careful reading of sources is needed to find matches. Relevant text extracts from matching sources or pointers to these are stored in the table adjacent with the sentence.
4. If a sentence does not match any of the cited sources, the next step is to do a plagiarism search in the Web. Systematically varied queries using basic search tools such as Google and Wikipedia or special services for plagiarism checking can be used.
5. If a substantial share of the article's text still lacks matching sources, it is reasonable to search for them outside the public Web. This may include textbooks used in the school, materials of the school's learning environment or other materials available to the students. The success in the plagiarism search outside the public Web is heavily dependent on contextual and situational factors. Data collected about the task performance process, for example by observation and interviews, help in focusing search efforts.

The comparison of a sentence and sources is based on the analysis of literal and semantic content. Finding a match is a simple task in case of copy-pastes but becomes more challenging when the author extensively uses his or her own terminology and synthesises across sources. Another problem is that similar information may be available through several sources. The analyst must assess if the sentence could be formulated using one source only. An opposite problem is that the sentence can only be composed by combining information from several sources. Occasionally it occurs that no matching sources are found for a sentence. The percentage of sentences for which sources are found is named here *source recall*.

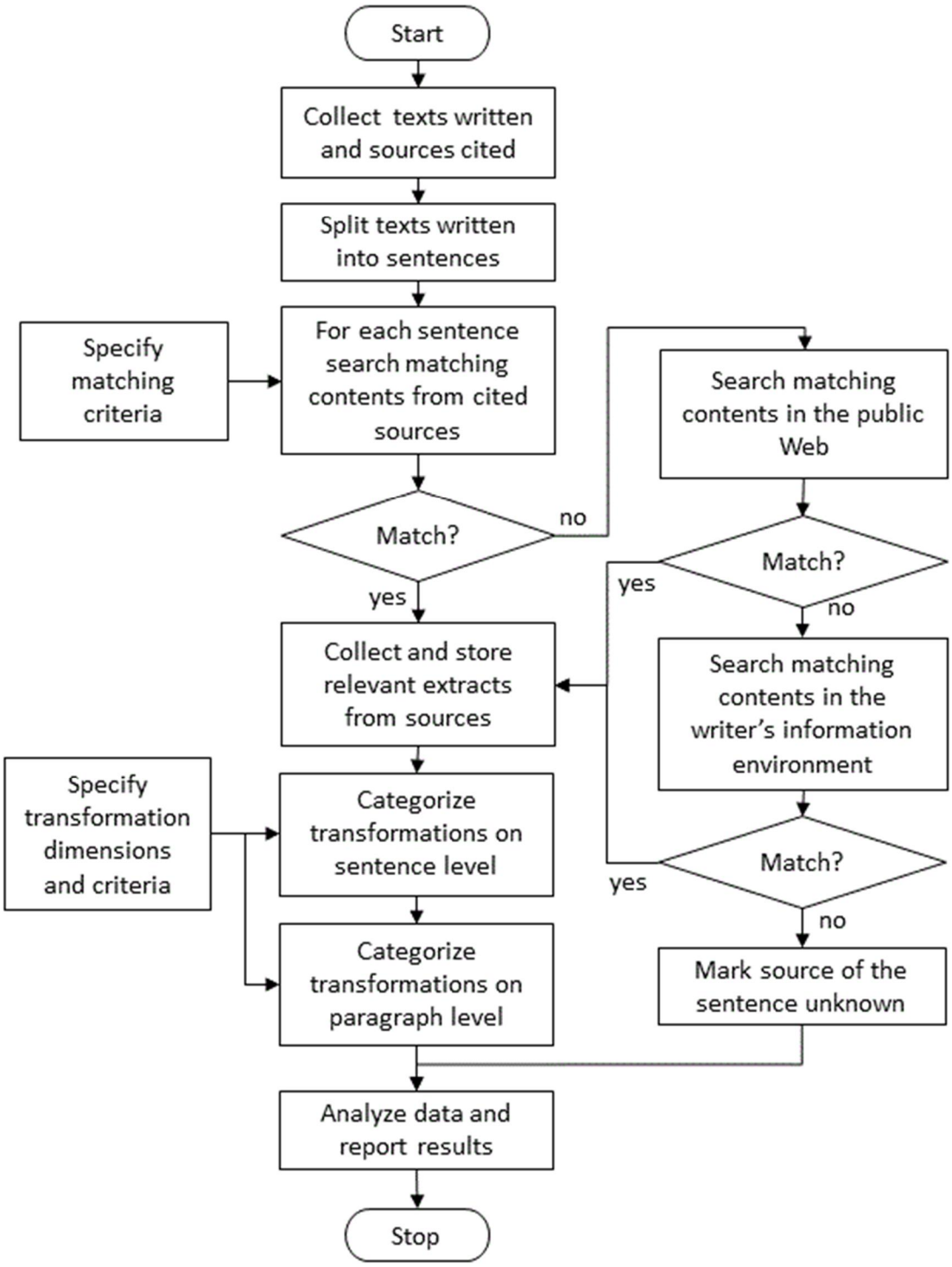

Figure 1: Flowchart of the proposed method

Text transformation categories

The categories for the text transformations developed in the pilot study combined ideas applied in plagiarism and reading-to-write studies (Sormunen and Lehtiö 2011). The framework was very

simple and merged two variables, the degree of paraphrasing and the degree of synthesis, into one-dimensional scale. The framework proposed here consists of six dimensions (variables). Four variables are related to the use of sources, one to the type of sentence written and one to the type of sources used. The dimensions are the following (sources which informed the development of categories are mentioned in parentheses):

1. The degree of paraphrasing (McGregor and Streitenberger 2004; Wiley and Voss 1999)
2. The degree of synthesis (Mateos and Sole 2009; Segev-Miller 2004; Spivey 1997)
3. Credibility in building arguments on the basis of sources (inspired by Hart 1998)
4. Accuracy of citing (plagiarism: McGregor and Streitenberger 2004; other categories derived from the data)
5. Statement type (inspired by Hart 1998: 89-93)
6. Source type (any typology for formal information sources applies)

The degree of paraphrasing indicates how much the writer uses his or her own words in constructing meaning for the text (categories: *copy-paste, near copy-paste, paraphrased, own text, and 'not applicable'*). Copy-pastes are exact copies and near copy-pastes slightly edited copies of the source text (e.g., word order changed, some words added or removed). *Paraphrasing* requires a major change beyond technical editing in expressing the content of the source. *Own text* refers to sentences which are comments or remarks made by the writer. Informative sentences likely to be derived from unknown sources are assigned to category *not applicable*.

The degree of synthesis measures the extent to which the writer connects bits of information from different parts of a single source or from multiple sources (categories: *sentence, paragraph, source, multiple sources, combined with own text, not applicable*). *Sentence, paragraph, source* means that a written sentence contains or summarizes information from a single sentence, a single text paragraph or several paragraphs of a single source. *Multiple sources* means that information is derived from two or more sources. *Combined with own text* indicates that the sentence is partially based on sources and partially on the writer's own knowledge or views.

In the study of credibility dimension the aim is to assess how solidly arguments expressed in a sentence base on the evidence presented in the sources used (categories: *credible, ambiguous, error, weak source, not applicable*). The goal is not to assess the *truth* or *correctness* of information. *Ambiguous* means that the writer fails to represent the piece of information derived from the source clearly, thus increasing the risk that the reader acquires biased or faulty facts or interpretations. *Error* denotes a definite logical error in interpreting the source and weak source indicates that the quality of the source is dubious.

Accuracy of citing deals with the punctuality of linking the written sentence to sources used (*categories: sentence, paragraph, article, other, plagiarism, cheating, not applicable*). Categories *sentence, paragraph, article, and other* express how close to the sentence the corresponding in-text citation is located. Category *plagiarism* denotes that one or more non-cited sources were found for the sentence. In the case of *cheating* the writer has plagiarised a source but given a reference to a inapplicable source.

Dimension *Statement type* introduces a simple idea to apply argumentation analysis in categorizing sentences written by students. We applied the typology of claims introduced by Hart (1998: 89-93) and propose the following categories: *claims of concept* (definitions), *claims of fact* (*objective* facts), *claims of interpretation* (*subjective* facts), *claims of value, technical comments*. *Claims of fact* refers to statements that can be proven true or false (e.g. 'Helsinki is the home base of many

leading Finnish export industries.'). *Claims of interpretation* are proposals as to what facts mean (e.g. 'Helsinki has a major role in Finland's economy'). *Claims of value* are judgements about the worth of something.

Source type expands the view on materials exploited by the writer. The relevance of categories depends on the context of writing. For example, one may apply main categories of printed and Web sources and their subcategories.

Case study

Research questions

The goal of the empirical part of the study is to demonstrate how the method proposed can be applied to the analysis of text transformations. The main research question is: How do the students of an upper secondary school use information derived from sources in writing an article as a collaborative learning assignment?

We look for answers to the main question from several viewpoints by formulating the following sub-questions:

Research question 1. To what extent do texts composed by students in collaborative source-based writing contain sentences

- paraphrasing beyond copy-pasting?
- summarising within and synthesising across sources?
- plagiarising sources?
- building arguments credibly on sources?
- expressing claims of facts and interpretation?

Research question 2. Is the use of printed and web sources similar or different in terms of

- copy-pasting?
- plagiarism?
- summarising within and synthesising across sources?
- building arguments credibly on sources?
- expressing claims of facts and interpretation?

Research question 1 demonstrates how the method can be used to make a descriptive study on the different aspects of texts written by students. Research question 2 expands the analysis to the relationship of the text aspects and source types.

Case courses

Data were collected from two eight-week courses in an upper secondary school in the city of Tampere, Finland, during the spring term 2011. Thirty students organized into ten groups (three members in each) completed a course in Finnish literature. Twenty-eight students organized into

seven groups completed a course in Finnish history: two three-member, three four-member and two five-member groups. The members were allocated into groups randomly by lot.

On the literature course, the task was to write an article for the Finnish edition of Wikipedia. The history course used a dedicated school wiki as the writing forum. On both courses, the assignment was designed to follow Wikipedia's conventions and requirements for authors. The student groups selected a topic for their article from a list prepared by the teacher.

On the literature course each assignment was about a classic Finnish novel. The students were required to read the novel first and then write a personal literary essay about it before the group work started. The teams were required to write about the novel, about the author and his or her works overall, about the reception of the novel in its time, etc.

On the history course, the teacher had prepared topics dealing with Finnish history from the Civil War to the beginning of the Winter War (1918-1939). The topics were quite extensive: The Civil War (1918), a dispute over the Finnish constitution (1918-19), economic development, the role of the left wing, the role of the right wing and foreign policy. The articles on the last four topics were intended to cover the period 1918-39. For each topic, the teacher had listed sub-topics to help students comprehend what the article should contain.

The total time reserved for the assignment was thirteen days in the history class and thirty days in the literature class (including time for reading the novel and preparing the literary essay). On both courses the assignment was introduced, written guidelines were distributed, groups formed, and topics for the articles selected at the first meeting. The second meeting was a visit to the nearby city library. One thirty-minute lesson was devoted to the library collections and services and another lesson to searching in the internet. The librarian was informed of the topics selected and had collected materials from the library collection for the students to look at.

After the visit to the library the students worked the next five (in the history course four) lessons in the computer class to search for information, to select and read sources found and to write text for the articles under the teacher's supervision. On the history course a substitute teacher was supervising the class for two lessons instead of the regular teacher.

Data collection

This article only reports findings on the text analysis. The data from the main study, however, include a rich material consisting, for instance, of student interviews and questionnaires. In the process of text analysis the interview material was occasionally consulted to aid in finding information sources.

Two second-year Master's degree students were hired as research assistants. One of them was studying Finnish literature and the other history. We aimed to ensure that both analysts had enough background knowledge in the subject area of articles and sources for consistent and informed decisions.

All articles written by the students were split into sentences and stored into Microsoft Word tables. For each sentence, the sources used were searched starting from the closest text citations and expanding to plagiarism tests in the Web. If a substantial share of article sentences still lacked identified sources (> 10%) we checked the textbooks available to the students and the materials

mentioned by students in the interviews. Relevant extracts from the sources identified were stored in the table of sentences.

Table 1: Consistency of coding in a two-phase test

Test set		Degree of paraphrasing	Degree of synthesis	Accuracy of citing	Credibility of arguments	Type of claims	Type of sources	Overall
Consistency test 1	Article H1	85 %	68%	85 %	85 %	76 %	74 %	79 %
	Article L7	82 %	78 %	74 %	86 %	86 %	90 %	82 %
Consistency test 2	Article H2	94 %	90 %	93 %	81 %	90 %	95 %	91 %
	Article L1	94 %	96 %	98 %	91 %	98 %	96 %	95 %

Plagiarism checking was mainly done by Google and Wikipedia searches. To avoid the problems of variation in text transformations we selected up to four “best” words from each sentence as keyword candidates for Wikipedia searches. Queries were made with all permutations of two keywords (maximum of six queries if necessary). A similar procedure was applied in Google searches but a set of five keyword candidates was used first. The queries were made using all keywords and combinations of four keywords (maximum five attempts). In each search result, a set of twenty first hits was checked. If more than ten per cent of sentences were still missing a source, the plagiarism search was expanded to printed materials mentioned in the student group interviews.

Data analysis

The research assistants familiarised themselves with the task by applying the codes to one article in their own area of expertise (literature or history). On the basis of the problems encountered the codes and coding guidelines were revised. Next, both research assistants independently coded a set of two articles (literature and history) as the first consistency test. The differences in codes were analysed and the definitions and interpretations of some codes were discussed and revised. The level of consistency was again checked by coding a new set of two articles. The overall consistency in the first round was 79-82% and rose to 91-95%in the final round (Table 1).

After the consistency tests each research assistant coded the data of her/his course alone. To balance the workload one of the history articles was analysed by the literature expert. The summary of data analysed is presented in Table 2. The volume of texts analysed in the set of literature articles was smaller because we excluded sections based on the direct literary analysis of the novel (description of the plot and characters). The third column of the table indicates in how many sentences the analyst was unable to identify the source of the sentence. The overall share of sentences where the search for sources failed was about 6 per cent for history articles and about 11 per cent for literature articles. The coding data was first collected into Excel tables and after error-checks and pre-processing transferred into SPSS software. All variables to be examined were categorical and thus the data were organized into 2 x 2 or 2 x 3 contingency tables. The chi-squared (χ^2) measure was used to test the statistical significance of differences in the distributions of column and row frequencies. This has been applied to cross-tabulated categorical data (Reinhard 2006.) The χ^2 test can be used if $N > 40$, if fewer than 20% of the cells have an expected frequency of less than 5 and

if no cell has an expected frequency of less than 1 (Siegel and Castellan 1988: 123-124). We used $p<0.05$ as the critical limit for statistical significance.

Table 2: Summary of articles and sources used

Team No.	Sentences		Cited sources used			Plagiarised sources			All sources used			Cited sources not used
	Analysed	Source unknown	Web	Printed	Subtotal	Web	Printed	Subtotal	Web	Printed	Total	
History course												
H1	116	8	11	4	15	4	1	5	15	5	20	1
H2	125	11	10	4	14	1	0	1	11	4	15	0
H3	68	3	12	3	15	10	0	10	22	3	25	0
H4	74	2	7	1	8	1	0	1	8	1	9	0
H5	143	12	12	4	16	3	1	4	15	5	20	0
H6	59	1	5	1	6	3	1	4	8	2	10	11
H7	116	6	10	5	15	1	1	2	11	6	17	2
Average	100.1	6.1	9.6	3.1	12.7	3.3	0.6	3.9	12.9	3.7	16.6	2.0
STDev	32.6	4.4	2.6	1.6	4.0	3.2	0.5	3.1	4.9	1.8	5.7	4.0
Literature course												
L1	46	0	6	3	9	0	0	0	6	3	9	2
L2	17	1	0	4	4	1	1	2	1	5	6	0
L3	18	1	1	3	4	0	0	0	1	3	4	0
L4	69	17	0	6	6	2	1	3	2	7	9	0
L5	36	0	4	6	10	1	0	1	5	6	11	1
L6	41	9	0	5	5	0	0	0	0	5	5	0
L7	48	5	0	4	4	2	0	2	2	4	6	1
L8	30	0	2	4	6	6	1	7	8	5	13	0
L9	23	3	0	3	3	0	0	0	0	3	3	0
L10	24	1	3	2	5	2	0	2	5	2	7	2
Average	35.2	3.7	1.6	4.0	5.6	1.4	0.3	1.7	3.0	4.3	7.3	0.6
STDev	16.3	5.5	2.1	1.3	2.3	1.8	0.5	2.2	2.8	1.6	3.2	0.6

The Cramér coefficient (V) was used as the measure of association between column and row variables. The advantages of the coefficient are that it can be used to compare contingency tables of different sizes and based on different sample sizes (Siegel & Castellan 1988: 232). In principle, the values range from 0 to 1 but rarely achieve a value higher than 0.80. We used the following criteria adopted from Rea and Parker (1997) to evaluate the values of V: 0.10 = no relationship; 0.10 … <0.20 = weak association; 0.20 … <0.40 = moderate association (no higher values found in this study).

Findings

Research question 1: To what extent do students transform information from sources?

The characteristics of texts written by students in the history and literature classes are summarised in Table 3. The results are presented separately for both history and literature classes.

Table 3: Characteristics of sentences in articles written by students in a source-based writing assignment.
Differences related to assignment types and to type of sources used.

The aspect of source-based writing		**Class**			**Source type**		
		History (n=653)	**Literature (n= 292)**	**Total (n=945)**	**Printed sources (n=486)**	**Web sources (n=430)**	**Total (n=916)**
Degree of paraphrasing	copy-paste	9%	9%	9%	5%	14%	9%
	near copy-paste	42%	26%	37%	38%	38%	38%
	paraphrased	49%	65%	54%	57%	48%	53%
Total		100%	100%	100%	100%	100%	100%
		χ^2 (2)=23.5; p=0.000; V=0.158			χ^2 (2)=21.8; p=0.000;V=0.154		
Degree of synthesis	sen-sen	61%	53%	58%	60%	60%	60%
	summary	34%	37%	35%	36%	37%	36%
	synthesis	5%	10%	7%	4%	3%	4%
Total		100%	100%	100%	100%	100%	100%
		χ^2 (2)=10.8; p=0.004;V=0.107			χ^2 (2)=0.125; p=0.939;V=0.012		
Accuracy of citing	close	65%	49%	60%	67%	51%	60%
	loose	19%	34%	24%	28%	20%	24%
	missing	16%	16%	16%	5%	29%	16%
Total		100%	100%	100%	100%	100%	100%
		χ^2 (2)=28.3; p=0.000; V=0.173			χ^2 (2)=97.7; p=0.000;V=0.327		
Credibility of arguments	no problem	87%	82%	85%	86%	84%	85%
	problem	13%	18%	15%	14%	16%	15%
Total		100 %	100 %	100 %	100 %	100 %	100 %
		χ^2 (2)=3.2; p=0.072;V=0.072			χ^2 (2)=0.747; p=0.388;V=0.029		
Type of claims	fact	67 %	45 %	60 %	55 %	65 %	60 %
	interpretation	31 %	52 %	37 %	42 %	33 %	38 %
	other	2 %	3 %	3 %	3 %	2 %	3%
Total		100 %	100 %	100 %	100 %	100 %	100 %
		χ^2 (2)=40.0; p=0.000;V=0.206			χ^2 (2)=8.3; p=0.015;V=0.095		

Degree of paraphrasing. Overall, slightly more than a half of the sentences were written in paraphrased form and the share of copy-pasted text was 9%. Copy-pasting was equally common in both assignment groups but paraphrasing was more typical in the literature assignments (65%) than in the history ones (49%). The difference in the distributions between groups was statistically significant ($\chi^2(2)=23.5$; p=0.000) but the degree of association was weak (V = 0.158).

Degree of synthesis. The results showed that most sentences written (58%) were derived from a single sentence in the source. About one third of sentences (35%) summarised the content of a single source. Seven per cent of sentences synthesise contents from two or more sources. A small but statistically significant difference was observed between the groups ($\chi^2(2)=10.8$; p=0.004). The students on the literature course showed a tendency to summarise and synthesising more actively (weak association, $V = 0.107$).

Accuracy of citing. The overall result was that a text citation was assigned to 60% of sentences at the level of sentence or text paragraph. In about a quarter of sentences (24%) the source was in the list of references but no in-text citation was used to explicitly link the sentence to the sources. In 16% of sentences the source was plagiarised. The history group was more meticulous in marking text citations ($\chi^2(2)=28.3$; p=0.000) but the degree of association was weak ($V = 0.173$) . The most interesting result was that plagiarism was equally common in both groups (16%).

Credibility of arguments. The results showed that the students built their arguments well on the sources. In 85% of sentences, we could not find even minor problems in students' ways of interpreting the content of sources. The difference measured between courses was not statistically significant.

Type of claims. Overall, a high percentage of text (60%) reproduced facts from sources. More than a third of sentences (37%) contained source-based interpretations of facts. Only three per cent of sentences contained a definition or a claim of value. The comparison of group distributions suggested that the history group focused more on facts (67% vs. 31%) and the literature group on the interpretation of facts (45% vs. 52%). The difference measured was statistically significant ($\chi^2(2)=40.0$; p=0.000) and the degree of association moderate ($V = 0.206$).

RQ2: Is the use of information similar in printed and web sources?

The characteristics of texts written on the basis of printed and web sources are presented in Table 3 above. Twenty-nine sentences synthesising information from both printed and web sources were excluded from the basic data set reducing it to 916 sentences. The exclusion especially affected the distributions in the degree of synthesis.

Degree of paraphrasing. The figures reveal the tendency towards direct copy-pasting in the use of Web sources (14% vs. 5%) and towards paraphrasing in the use of printed sources (57% vs. 48%). Slight technical transformations were equally common (38%) in the use of both source types. The differences measured in copy-pasting and paraphrasing were statistically significant ($\chi^2(2)=21.8$; p=0.000). The result suggests that students exploited the technical ease of copying web sources (weak association, $V = 0.154$).

Degree of synthesis. We could not reject the null hypothesis concerning the degree of synthesis in students' articles ($\chi^2(2)=0.125$; p=0.939). The students seemed to summarise and synthesise information at a similar rate both from printed and web sources.

Accuracy of citing. The results corroborate the generally shared observation that plagiarism is associated especially with the use of web sources. Twenty-nine per cent of sentences making use of Web sources were products of plagiarism while this remained at the level of 5% in texts relying on printed sources ($\chi^2(2)=97.7$; p=0.000). The degree of association was moderate ($V = 0.327$).

If we exclude plagiarised sentences, the type of source did not seem to affect the accuracy of citing. Text citations were close to the sentence in 71-72% and loose in 28-29% of cases both for printed and web sources ($\chi^2(2)=0.116$; p=0.733).

Credibility of arguments. No difference was observed between source types in the credibility of building arguments ($\chi^2(1)=0.747$; p=0.388).

Type of claims. The analysis of sentence types written revealed a tendency to collect more facts from the Web (65% vs. 55%) while interpretations of facts were more commonly derived from printed sources (42% vs. 33%). Other types of sentences had a minor role and do not affect the general trend. The difference measured was statistically significant ($\chi^2(2)=8.3$; $p=0.015$) but the difference observed was hardly of practical importance ($V = 0.095$, no relationship).

Discussion of empirical findings

The findings of the present study demonstrate that the proposed method enables a detailed analysis of the use of sources in writing assignments. We discuss some of the findings to explicate our contribution.

Research question 1, The answers to the first research question emphasize that students used sources and transformed information differently in differently profiled assignments (see subsection *Case courses*). However, we cannot draw any conclusion on which of the differences in the assignment characteristics caused the differences in the written outcomes. On the other hand, the results also suggest that some specific information practices were similar across the courses. All the students were from the same school, the groups were large and the courses were compulsory for all students (no pre-selection). Thus shared practices could be associated with the particular school and schooling there.

Two interesting similarities were observed in the articles written on the two courses. The groups composed copy-pasted (9%) and plagiarised (16%) sentences equally often. The finding suggest that copy-pasting and plagiarism are deeply integrated into the practice of schooling (*school culture*) and these behaviours are not sensitive to minor changes in the way source-based writing assignments are designed and introduced (cf. Limberg *et al.*2008).

In other practices of source-based writing we identified differences between the courses. In the literature assignment, the students were more active in paraphrasing, in summarising and synthesising and in writing interpretative sentences beyond reproducing facts. The students of the history class were behind in all three aspects of source-based writing but they cited sources more carefully.

We can only present some preliminary hypotheses on the reasons why the articles on the literature course were more advanced in terms of certain evaluation criteria. The design of the literature assignment was more focused (the classic novel anchored the topic for each student team). The personal essay prepared students for searching information and writing the article. The structure and content of the required end-product were more explicitly specified (a Wikipedia article of a

particular type), and the progress of groups was monitored more intensively (checkpoints, the regular teacher was present at all lessons). Earlier research suggests that keeping the contextual aspects of the assignment simple and fixed seems to help students to focus on the contents of the assignment and achieve better outcomes (cf. Limberg *et al.* 2008; Hongisto & Sormunen 2010).

Research question 2. The comparison of information use derived from printed and Web sources suggests that 1) copy-pasting instead of paraphrasing, 2) plagiarism instead of acknowledging sources and 3) sentences reproducing facts instead of interpretations are more common when using Web sources. However, the findings are not conclusive since we did not eliminate the effect of assignment type (students in the literature class used more printed sources). No difference was found in the degree of synthesis or in the credibility of building arguments on sources. The findings exemplify the commonly held view that the Web has an obvious role in inappropriate practices of source-based writing (cf. e.g. McGregor and Williamson 2005; Purdy 2010).

Discussion

The method proposed was developed for the analysis of texts composed by students in source-based writing assignments. It is possible that some parts of it could be automated and applied by teachers in the assessment of students' texts. It is also possible that the area of application could be extended to other kinds of texts composed in source-based writing. However, we focus here on justifying the method only in the original application context.

The *appropriateness* of a method for the use intended can be justified, for example, by showing that it is possible to study new types of research questions or address old research questions in a more fruitful way. The review of related research revealed that studies on information use have not penetrated to the level of text transformations. Researchers in reading-to-write studies have applied similar text analysis but only in experimental settings and the focus has been different (e.g. Rouet 2006; Wiley and Voss 1999). The findings of the case study demonstrated the nature of empirical questions that would be beyond the reach of commonly used data collection approaches such as observations, interviews or questionnaires, but could be tackled by the method developed.

The *validity* of the method means that it is based on an established interpretation of essential variables in the phenomenon observed. The core variables and their operationalization were introduced in the framework of text transformation categories. We adopted most of the categories from reading-to-write studies (e.g. Wiley and Voss 1999), plagiarism studies (e.g. McGregor and Streitenberger 2004) and argument analysis (Hart 1998) although clarification and redefinition were required. The framework of transformation categories is open to extensions.

The *reliability* of the method is obviously sensitive to the level of source recall (the percentage of sentences for which sources were successfully found) and to the consistency of coding. In the case study, source recall was, on average, 94% (history course) and 89% (literature course). These figures are quite high and indicate that the findings would only have changed slightly if sources had been found for all sentences. The 6% or 11% of the used sources that could not be found is a possible warning sign of potentially poor citing practices. Extensive paraphrasing may also decrease source recall. However, paraphrasing should have a minor effect on source recall. If the reader of a text cannot see the content connection to the source, the writer has deviated from the good practice of source-based writing. The other aspect of reliability, consistency of coding was high in the case study: about 79-82% in the first test and 91-95% in the second test.

The *efficiency* of the procedure is a relative issue. The case study indicated that the method could be used successfully by exploiting the limited resources of a research project. Two research assistants each worked for twelve weeks for twenty hours a week The project paid for 480 hours in total (three months). We consider this a reasonable investment in the data.

The limitations of the method include its exclusive focus on text analysis. For a more thorough understanding of the use of sources it would be beneficial to include e.g. interviews with students which could shed light on specific behaviours detected through the text analysis. Interviews would also afford an opportunity to verify the researchers' selection of sources. A further extension of the method could include asking students to record the keywords they used in their searches, or automatically recording their searches.

Conclusions

We have introduced a novel method for the analysis of text transformations and information use in source-based writing. The appropriateness, validity, reliability and efficiency of the method were discussed above. We argue that the method opens up new possibilities for studying the core aspects of information use by focusing on text transformation from sources to written texts. The case study demonstrates that the method can be used productively in examining not only copy-pasting and plagiarism but also higher level text transformations such as the synthesis of information and quality of arguments.

No single method solves the problems of studying a complex phenomenon such as information use in source-based writing. One direction is to develop methods for more ambitious studies that codes for evidence of a critical view of the information from different sources, for example, in comparing and contrasting two sources, or a cause and effect from two sources.

The study of the end-product calls for approaches to study the process of source-based writing. In the case study, we collected a rich set of interview and survey data and look forward to combining different data sets to further elaborate information use in school assignments.

Acknowledgements

The study was part of the Know-Id project and the first author's sabbatical project funded by the Academy of Finland (grants no. 132341 and no. 136401). The authors thank the teachers of the case courses and the *Tieto haltuun* project in the City of Tampere for cooperation in data collection. We are grateful to Leeni Lehtiö and Teemu Mikkonen, who took care of the data collection during the case courses.

About the authors

Eero Sormunen is a Professor in the School of Information Sciences, University of Tampere, Finland. Sormunen received his Master of Science (Electrical Engineering) in 1978 from the Tampere University of Technology and his PhD (Information Studies) in 2000 from the University of Tampere, Finland. He is the leader of the project Informed Learning and Social Media in School and beyond funded by the Academy of Finland for years 2010-2013. Sormunen can be contacted at: eero.sormunen@uta.fi.
Jannica Heinström is a Senior Lecturer in Information Studies at Åbo Akademi University, Finland. Currently she holds a Senior Research Fellowship at the Institute for Advanced Social

Research, University of Tampere (2012–2013). Jannica received her Master's degree in Psychology (1994) and her PhD in Information Studies (2002) from Åbo Akademi University. She is working as a researcher in the project Informed Learning and Social Media in School and beyond funded by the Academy of Finland. Jannica can be contacted at: Jannica.Heinstrom@uta.fi.

Leena Romu worked as a research assistant and contributed to the case study in the development of text transformation categories and coding. She currently works on her PhD thesis in the School of Language, Translation and Literary Studies at the University of Tampere. She can be contacted at Leena.Romu@uta.fi

Risto Turunen worked as a research assistant and contributed to the case study in the development of text transformation categories and coding. He recently completed his Master's degree in the School of Social Sciences and Humanities at the University of Tampere. She can be contacted at Risto.Turunen@uta.fi

References

- Alexandersson, M. & Limberg, L. (2003). Constructing meaning through information artefacts. *New Review of Information Behaviour Research*, **4**(1), 17–30.
- Brookes, B.C. (1980). The foundations of information science. Part I: philosophical aspects. *Journal of Information Science*, **2**(3-4), 125-133.
- Cole, C. (1997). Information as process: the difference between corroborating evidence and "information" in humanistic research domains. *Information Processing & Management*, **33**(1), 55-67.
- Davis-Lenski, S. & Johns, J.L. (1997). Patterns of reading-to-write. *Reading Research and Instruction*, **37**(1), 15-38.
- Hart, C. (1998). *Doing a literature review. Releasing the social science research imagination*. London: Sage Publications.
- Heinström, J. (2002). *Fast surfers, broad scanners and deep divers - personality and information seeking behaviour*. Åbo (Turku), Finland: Åbo Akademi University Press. (Doctoral dissertation).
- Heinström, J. (2006). Fast surfing for availability or deep diving into quality – motivation and information seeking among middle and high school students. *Information Research*, **11**(4), paper 433. Retrieved 4 May 2011 from http://informationr.net/ir/11-4/paper265.html. (Archived by WebCite® at http://www.webcitation.org/6BSjSXs2H)
- Hongisto, H. & Sormunen, E. (2010). The challenges of the first research paper – observing students and the teacher in the secondary school classroom. In A. Lloyd & S. Talja (Eds.) *Practising information literacy: bringing theories of learning, practice and information literacy together*. (pp. 95-120). Wagga Wagga: Centre for Information Studies. Retrieved 21 September, 2012 from https://www12.uta.fi/blogs/know-id/files/2010/05/Hongisto_Sormunen_PIL2010.pdf. (Archived by WebCite® at http://www.webcitation.org/6BSjixPrg)
- Kari, J. (2007). Conceptualizing the personal outcomes of information. *Information Research*, **12**(2) paper 292. Retrieved 12 December, 2011 from http://InformationR.net/ir/12-2/paper292.html. (Archived by WebCite® at http://www.webcitation.org/6BSjvPlLV)
- Kari, J. (2010). Diversity in the conceptions of information use. *Information Research*, **15**(3), colis709. Retrieved 11 January 2012 from http://InformationR.net/ir/15-3/colis7/colis709.html (Archived by WebCite® at http://www.webcitation.org/6BSk2Eah6)
- Kiili, C., Laurinen, L. & Marttunen, M. (2008). Students evaluating Internet sources: from versatile evaluators to uncritical readers. *Journal of the Educational Computing Research*, **39**(1), 75-95.

- Kiili, C., Laurinen, L., Marttunen, M., & Leu, D. J. (in press). Working on understanding during collaborative online reading. *Journal of Literacy Research*, [planned vol. **44**, no. 4].
- Kirk, J. (2002). Theorising information use: managers and their work. Unpublished doctoral dissertation, University of Technology, Sydney, Australia. Retrieved 11 January 2012 from http://epress.lib.uts.edu.au/dspace/bitstream/handle/2100/309/02whole.pdf?sequence=2 (Archived by WebCite® at http://www.webcitation.org/6BSkLCz88)
- Limberg, L. (1999). Experiencing information seeking and learning. *Information Research*, **5**(1), paper 68. Retrieved 12 July 2011 from http://informationr.net/ir/5-1/paper68.html. (Archived by WebCite® at http://www.webcitation.org/6BSkSyACr)
- Limberg, L., Alexandersson, M., Lantz-Andersson, A. & Folkesson, L. (2008). What matters? Shaping meaningful learning through teaching information literacy. *Libri*, **58**(2), 82–91. Retrieved 16 October, 2012 from http://www.librijournal.org/pdf/2008-2pp82-91.pdf (Archived by WebCite® at http://www.webcitation.org/6BSkcmYrQ)
- Mateos, M. & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education*, **24**(4), 435-451.
- Maybee, C. (2006). Undergraduate perceptions of information use: the basis for creating user-centered student information literacy instruction. *The Journal of Academic Librarianship*, **32**(1), 79-85.
- McGregor, J. & Streitenberger, D. (2004). Do scribes learn? Copying and information use. In M. K. Chelton and C. Cool (Eds.), *Youth information-seeking behavior: theories, models and issues*. (pp. 95-118). Lanham, MD: Scarecrow Press.
- McGregor, J. & K. Williamson. (2005). Appropriate use information at the secondary school level: Understanding and avoiding plagiarism. *Library and Information Science Research*, **27**(4), 496-512.
- Newell, A. (1969). Heuristic programming: ill-structured problems. In J. Arofonsky, (Ed.). *Progress in operations research*, Vol. III, (pp. 360-414). New York, NY: John Wiley & Sons.
- Purdy, J.P. (2010). Wikipedia is good for you? In C. Lowe & P. Zemliansky, *Writing spaces: readings on writings* - Vol. 1. Anderson, SC: Parlor Press. Retrieved 12 July 2011 from http://wac.colostate.edu/books/writingspaces1/purdy--wikipedia-is-good-for-you.pdf. (Archived by WebCite® at http://www.webcitation.org/6BSoCnK52)
- Rea, L.M. & Parker, R.A. (1997). *Designing and conducting survey research: a comprehensive guide*. San Francisco, CA: Jossey-Bass.
- Reinard, J.C. (2006). Nonparametric tests for categorical variables. In *Communication research statistics*. (pp. 249-281). Thousand Oaks, CA: Sage Publications.
- Rouet, J-F. (2006). *The skills of document use. From text comprehension to Web-based learning*. Mahwah, NJ: Lawrence Erlbaum.
- Savolainen, R. (2009a). Epistemic work and knowing in practice as conceptualizations of information use. *Information Research*, **14**(1) paper 392. Retrieved 12 December, 2011 from http://InformationR.net/ir/14-1/paper392.html (Archived by WebCite® at http://www.webcitation.org/6BSo8XCUN)
- Savolainen, R. (2009b). Information use and information processing. Comparison of conceptualizations. *Journal of Documentation*, **65**(2), 187-207.
- Segev-Miller, R. (2004). Writing from sources: the effect of explicit instruction on college students' processes and products. *Educational Studies in Language and Literature*, **4**, 5–33.
- Siegel, S. & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences*. New York, NY: McGrawHill.
- Sormunen, E. & Lehtiö, L. (2011). Authoring Wikipedia articles as an information literacy assignment – copy-pasting or expressing new understanding in one's own words?

Information Research, **6**(4) paper 503. Retrieved 25 April 2012 from http://InformationR.net/ir/16-4/paper503.html (Archived by WebCite® at http://www.webcitation.org/6BSo3X5ME).

- Spivey, N.N. (1997). *The constructivist metaphor. Reading, writing, and the making of meaning*. San Diego, CA: Academic Press.
- Tuominen, K., & Savolainen, R. (1997). A social constructionist approach to the study of information use as discursive action. In: P. Vakkari, R. Savolainen, & B. Dervin (Eds.), *Information seeking in context: proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*. (pp. 81-96). London: Taylor Graham.
- Todd, R.J. (1999). Utilization of heroin information by adolescent girls in Australia: a cognitive analysis. *Journal of the American Society for Information Science*, **50**(1), 10-23.
- Todd, R.J. (2006). From information to knowledge: charting and measuring changes in students' knowledge of a curriculum topic. *Information Research*, **11**(4), paper 264. Retrieved 12 Dec 2011 from http://InformationR.net/ir/11-4/paper264.html (Archived by WebCite® at http://www.webcitation.org/6BSnv1vG8).
- Vakkari, P. (1997). Information seeking in context: a challenging meta-theory. In: P. Vakkari, R. Savolainen & B. Dervin, (Eds.), *Information seeking in context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*. (pp. 451-646). London: Taylor Graham.
- Vakkari, P. (2008). Trends and approaches in information behaviour research. *Information Research*, **13**(4), paper 361. Retrieved 05 January 2012 from http://InformationR.net/ir/13-4/paper361.html (Archived by WebCite® at http://www.webcitation.org/6BSnpdIEN).
- Wiley, J. & Voss, J.F. (1999). Constructing arguments from multiple sources: tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, **91**(2),301-311.
- Wilson, T.D. (2000). Human information behaviour. *Informing Science*, **3**(2). Retrieved 05 January 2012 from http://www.inform.nu/Articles/Vol3/v3n2p49-56.pdf (Archived by WebCite® at http://www.webcitation.org/6BSnkgbVO)

How to cite this paper

Sormunen, E., Heinström, J., Romu, L. & Turunen, R. (2012). "A method for the analysis of information use in source-based writing." *Information Research*, **17**(4), paper 535. [Available at http://InformationR.net/ir/17-4/paper535.html]