TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Camilla Magnusson

Text Visualization for Competitive Intelligence

Julkaisu 931 • Publication 931

Tampere 2010

Tampereen teknillinen yliopisto. Julkaisu 931
Tampere University of Technology. Publication 931

Camilla Magnusson

Text Visualization for Competitive Intelligence

Thesis for the degree of Doctor of Philosophy to be presented with due permission for public examination and criticism in Auditorium 125 at Tampere University of Technology – Pori, on the 19th of November 2010, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2010

ISBN 978-952-15-2466-0 (printed)
ISBN 978-952-15-2511-7 (PDF)
ISSN 1459-2045

ABSTRACT

The overreliance on quantitative, numeric information and the underuse of qualitative, textual information is a common weakness in strategic management. This affects particularly the practice of competitive intelligence, which aims to provide actionable information about the company-external environment for decision making. Yet, forward-looking and insightful information about the environment often comes in qualitative form, much of it as publicly available text documents, whereas public quantitative information often arrives too late to be useful for strategic management.

This thesis proposes that text visualization, and in particular, a method called collocational networks, could increase the use of textual information in competitive intelligence. Collocational networks are networks consisting of words that co-occur in a statistically significant way in a text, or in a collection of texts. They are particularly useful for discovering changes between sequences of texts of a similar nature, e.g. annual or quarterly reports.

In line with design science research practice, the proposed method is also evaluated for utility. The evaluation is carried out in two stages. First, collocational networks of quarterly report texts are compared to self-organizing maps created out of the financial figures of the same reports. This evaluation shows that changes in the collocational network of a company's quarterly report are followed by a change in the position of the company in the self-organizing map in the next quarter.

Second, a series of interviews with competitive intelligence practitioners are carried out, in which the interviewees are shown collocational networks created out of annual reports from telecommunications service companies during 2003-2008. This evaluation shows that the interviewees consider the networks to reflect actual developments within the industry. They also consider them to be a useful tool for discovering changes that may go unnoticed when reading the texts.

In summary, the evaluations suggest that by using collocational networks, competitive intelligence practitioners could easily have access to forward-looking, qualitative information about the company-external environment to strengthen strategic management practice.

Keywords: collocational networks, competitive intelligence, design science research, strategic management, text visualization

ACKNOWLEDGEMENTS

The origins of this PhD thesis lie in my Master's thesis, which I wrote in 2001-2002 within the Tekes GILTA project, which has been followed by the Tekes TITAN project. During the GILTA project, I got to know an inspiring group of researchers, some of whom I also collaborated with on the earlier research papers included in this thesis. For this fruitful collaboration I would like to thank my co-authors Dr. Antti Arppe, Prof. Barbro Back, Dr. Tomas Eklund, Prof. Hannu Vanharanta and Prof. Ari Visa.

In 2004, I began working full-time in industry and set the PhD thesis aside. During these years, I gained valuable insight into how my research could be linked to strategic management practice. When I returned to work on the thesis, Prof. Hannu Vanharanta became my supervisor, always encouraging and motivating me to continue with the work. I am very grateful for the time he has put into this process. Another important motivator at this later stage was the PhD student seminar in business information management held in Tampere during 2008-2009 by Prof. Samuli Pekkola and Prof. Hannu Kärkkäinen. The conversations with professors and students during (and outside) the seminar helped me see my work in a wider research context.

I am also grateful to those who have read and commented on my thesis or parts of it at different stages. One of them is Dr. Antti Arppe, whom I would like to thank for commenting on the text from a linguist’s point of view, and, not least, for getting me started on this research topic back in 2001. I would also like to thank Dr. Arto Siitonen for his comments on the methodology section and for always encouraging me to keep writing.

I would also like to express my gratitude to the reviewers of my thesis, Prof. Pekka Pihlanto and Dr. Charalampos (Harris) Makatsoris, for suggesting changes and additions that have greatly improved the thesis.

For the final parts of the work, I gratefully acknowledge the financial support of the Finnish Foundation for Economic Education, the KAUTE foundation, and the Finnish Cultural Foundation (Satakunta regional fund).

Finally, I am grateful for all the support I have received from friends, family and colleagues over the years. This work could never have been completed without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS

1. Magnusson Camilla & Hannu Vanharanta 2003. Visualizing sequences of texts using collocational networks. Petra Perner and Azriel Rosenfeld (Eds.) *Machine Learning and Data Mining in Pattern Recognition*. Lecture Notes in Artificial Intelligence 2734. Berlin: Springer. pp. 276-283.

2. Magnusson Camilla, Antti Arppe, Tomas Eklund, Barbro Back, Hannu Vanharanta & Ari Visa 2005. The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management*, 42, 561-574.

3. Magnusson Camilla 2010. Improving competitive analysis with temporal text visualization. *Marketing Intelligence & Planning, 28*, 571-581.

4. Magnusson Camilla, unpublished manuscript. A qualitative user evaluation of collocational topic networks as a text visualization method. Submitted to European Journal of Information Systems.

ABBREVIATIONS

IS	Information Systems
MI	Mutual Information
SOM	Self Organizing Map
VAS	Value Added Service

GLOSSARY

Artifact	An object made by a human, as opposed to an object occurring in nature. (alternative spelling: artefact)
Collocation	The co-occurrence of two or more words in a text or in a collection of texts.
Collocational network	A network consisting of words that co-occur in a text or in a collection of texts
Competitive intelligence	The gathering and analyzing of company-external information for decision making purposes.
Corpus linguistics	Branch of linguistics dealing with the analysis of collections of texts (i.e. corpora).
Data	Letters and figures with no intrinsic meaning. Cf. information.
Design science research	Research methodology for solving real-world problems through the creation and evaluation of an artifact.
Information	Data which has been given meaning in a human context.
Interpretivism	The view that knowledge is a matter of interpretation.
Knowledge base	The body of relevant research from which relevant knowledge can be drawn when designing a new artifact.
Ontology	A theory on the nature of the being of an object.
Positivism	The view that knowledge must be based on observable facts.
Self organizing map	An artificial neural network that can be used to represent large data sets.
Text linguistics	Branch of linguistics dealing with the analysis of texts.

1 INTRODUCTION

1.1 Background

This thesis addresses a bias that is common in strategic management practice: when trying to analyze their external environment, companies often overlook qualitative information in favour of quantitative information. Yet, qualitative information, much of it available in public text documents, often contains more forward-looking and insightful information concerning the environment than quantitative information. As Mintzberg (1994, 258) puts it: “a bias toward the quantitative can allow the economic to displace the social and the financial to displace the creative”.

One reason for such a bias is that qualitative information is perceived as fuzzy and difficult to analyze in a systematic way. To ease the use of qualitative information, this thesis presents a method based in linguistics for visualizing sequences of texts produced in an organization’s external environment, for example by its competitors, partners or suppliers. The method allows users to track changes in how a topic is presented from one text to the following. This allows them to detect rising trends before they begin to appear in quantitative, financial information.

1.2 Main aim

This thesis introduces a simple text visualization method intended to improve and increase the use of qualitative material in competitive intelligence. The method consists of visualizing changes in sequences of texts published by a company's competitors, partners or other strategically significant organizations. The method is built and evaluated in line with the work of Hevner et al (2004), a central work in design science research. This method is suggested as an improvement to current competitive intelligence practices which suffer from the dominance of quantitative information in analysis.

Underlying the research is the notion that design science should take into account the dual nature of the research artefact, here the text visualization method, as both a physical object and as an intentional object. This means that the development and evaluation of an artifact should be grounded in the human context for which it is intended. This is particularly important for an artifact as fundamentally based in the context of language use as a text visualization tool.

1.3 Thesis structure

The thesis is constructed so that the following chapter (Chapter 2) presents the design science research methodology that is used. In Chapter 3, the context of competitive intelligence and its significance for strategic management is discussed. There, the issue of overreliance on quantitative information in decision making is also discussed. In Chapter 4, collocational networks are presented as a simple visualization method that allows intelligence practitioners to bring in qualitative information into their analysis. Then, the method is evaluated in two ways: first, a quantitative evaluation is made (Chapter 5). Here, changes in collocational networks made out of corporate quarterly reports are compared to changes in a self-organizing map created out of the financial figures published in the same reports. A second evaluation, of qualitative nature, is then made using interviews with competitive intelligence practitioners (Chapter 6). The final chapter contains conclusions and suggestions for further research that have arisen during this process. For a slightly different view, Figure 1 shows how the contents of this thesis correspond to the requirements of the design science research process as defined by Hevner (2007).

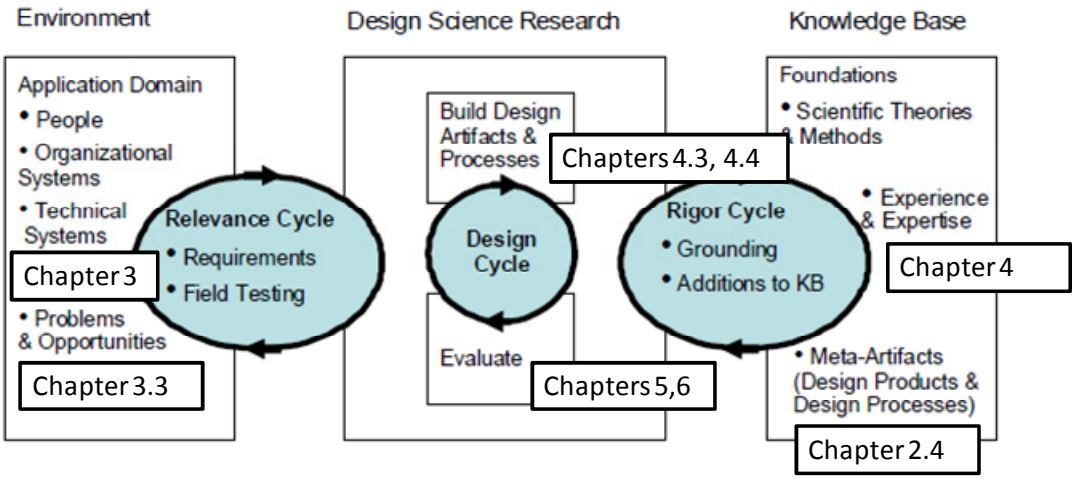

Figure 1. Structure of the thesis connected to the environment, design science research methodology, and knowledge base as defined by Hevner (2007)

Beginning on the left, the application domain of this thesis, competitive intelligence for strategic management purposes, is discussed in Chapter 3. Chapter 3.3 explicates the problems and opportunities of using qualitative information within this domain.

The main activities of the design cycle, the building of an artifact and evaluating it, are seen in the middle of this figure. The building process is discussed particularly in Chapters 4.3. and 4.4 and the evaluation is carried out in Chapters 5 and 6. Furthest to the right, the knowledge base containing text visualization and linguistic methods, is discussed in Chapter 4. The meta-artifact generated in this process, the view of the dual nature of the artifact, is presented in 2.4.

2 DESIGN SCIENCE METHODOLOGY

2.1 Introduction to design science

This thesis takes a design science approach to research. Most of current design science research cites Herbert A. Simon's seminal book, The Sciences of the Artificial (third revised edition 1996, originally published in 1969) as its origin. In his collection of papers, Simon sets out to define a science that complements natural science, a science of the artificial. Central to this science is the construction of an artificial object, an artifact, by the researcher, during the research process. The artifact is constructed to solve or alleviate a real-world problem.

Although design science research recently has created a wide interest within information systems research, the approach is by no means limited to information systems, but can be applied within other sciences as well. In fact, Simon believes that the process of design is relevant for schools of engineering, architecture, business, education, law and medicine. Within management science, van Aken (2004) in particular has recently been a champion of this approach.

Since the focus of this thesis lies at the intersection of management sciences and engineering sciences, both of these strands of research will be looked at more closely in this chapter. The literature review will, however, be mainly based on research done in information systems, as the body of design science research is currently more extensive within that science. Particularly the chapter on the design science research process (Chapter 2.3) will draw heavily on research conducted in information systems.

Within information systems research, the design science approach has gained much popularity during the 2000s. Although Simon's work was originally published in 1969 and some influential design science papers were published in the 1990s (Walls et al. 1992, Nunamaker et al. 1991), the popularity of design science research in information systems began to grow after the publication of Hevner et al. (2004) in high-ranking information systems journal MIS Quarterly. In their paper, Hevner et al. outline a framework for the iterative design science research process (see Figure 2, Chapter 2.3). They also propose seven guidelines for design science research, which will be discussed in more detail in Chapter 2.2.1. Indulska and Recker (2008), who have studied the methodologies used in papers presented at five major international information systems conferences, see a significant increase in conference papers dealing with design science after the publication of Hevner et al. (2004).

The publication of Hevner et al. (2004) not only increased the publication of research papers that follow a design science approach, but it also increased the interest in methodological issues related to design within the information systems research community. Some of these issues will be discussed further in Chapter 2.2.

According to Winter (2008), design science is currently the dominating information systems research paradigm in the German-speaking countries, and there are also a large number of design-oriented researchers in the Nordic countries, the Netherlands, Italy and France. Globally, design science research has received further attention within the information systems research community through the founding in 2006 of the annual DESRIST conferences, solely devoted to design science research. Design science has also been acknowledged as an important part of current information systems research through special issues in journals such as MIS Quarterly (2008), European Journal of Information Systems (2008) and the Scandinavian Journal of Information Systems (2007).

In management research, design science has received much less attention to date (2010). It has been championed particularly by van Aken (2004) who believes that management theory should develop into a design science in order to maintain its relevance for practitioners in the field. He calls for the development of "field-tested and grounded technological rules to be used as design exemplars of managerial problem solving", an approach mirroring design science in engineering.

2.2 Methodological background of design science research

From a philosophy of science point of view, the origins of the design science research approach lie in pragmatism, a school of thought developed by North American philosophers such as C. S. Peirce, William James and John Dewey around the late 19th and early 20th century. Pragmatism argues for a connection between knowledge and action. In a similar vein, Hevner et al. (2004) argue that whereas the goal of behavioural science is truth, the goal of design science is utility. They believe that truth and utility are inseparable, and thus connect design science to the tradition of scientific research, although they acknowledge that it operates in a way that differs from that of traditional research. The foundations of a design science research process lie in the construction of an artifact and the evaluation of its utility.

Goldkuhl (2008) distinguishes between three main types of pragmatism. In all of these, there is a connection between knowledge and action, but the form of connection is different in each. This thesis aims at full pragmatism, as called for by Goldkuhl in information systems research. Full pragmatism means that the three kinds of pragmatism all occur in the same work. The three kinds, and the way they are actualized in this thesis, are:

1. Referential pragmatism - producing knowledge that is useful and applicable for action. The artifact that is developed in this thesis,

collocational topic networks, is intended to be used as a tool in competitive intelligence practice.

2. Functional pragmatism - knowledge about action, describing the world in action-oriented ways. In this thesis, the literature review on competitive intelligence and strategic management as well as the evaluation interviews attempt to shed light on how intelligence activities are carried out in corporations.
3. Methodological pragmatism - producing knowledge through action. In a design science thesis, methodological pragmatism is essential. The creation and, in particular, the evaluation of the artifact produce information both about the artifact, collocational topic networks, and the context of competitive intelligence in which the artifact is intended to be used.

2.2.1 Positivist versus interpretivist approaches in design science

Much of design science in information systems carries a positivist undertone. The roots of positivism lie in Auguste Comte's argument that theories must be based on observed facts, a view that has shaped much of Western natural science practice. Simon (1996) compares design science with natural science and aims to define a science that has the same credibility and status that natural science.

Hevner et al. (2004), coming from an information systems background, do not speak of natural science but rather of behavioral science as the complement to design science. Their paper can, on some levels, be seen to encourage a more interpretivist approach to design science. For example, their proposed research guidelines are based on principles for interpretive field studies defined by Klein and Myers (1999). However, their framework for the research process bears a strong resemblance to the positivist ideals of natural science. Their outline of the design science research process as producing an artifact to solve a problem and then evaluating whether it works, resembles the prototypical natural science research process: producing a hypothesis to explain a naturally occurring

phenomenon and then conducting experiments to confirm or disconfirm the hypothesis.

Interpretivist research within information systems, is, however, not uncommon. In the early 1990s, Walsham (1995) noted a growing interest in interpretivist case research within the field and called for more researchers to include human interpretations and meanings in their studies of information systems. Ten years later, interpretivist research had become a well-established part of information systems research (Walsham 2006). The recent growth of design science research has also produced research that explicates its interpretivist stance on design. Niehaves (2007) argues for an interpretivist point of view of the design science research process, based on the seven guidelines for design science research laid out by Hevner et al. (2004), and on the set of principles for conducting interpretive field studies in information systems laid out by Klein and Myers (1999). Table 1 below contains an overview of both the guidelines of Hevner et al. and of Niehaves' commentary on five of them. It also shows how the guidelines have been applied to the design science research process that constitutes this thesis.

Guideline (Hevner et al. 2004)	**Hevner et al.'s point of view**	**Interpretivist commentary (Niehaves 2007)**	**Application of the guideline in this thesis**
1.Design as an Artifact	Design science must produce a viable artifact	(no additions to this guideline)	A text visualization method for competitive intelligence, collocational topic networks, is produced.
2.Problem Relevance	The object is to produce solutions to important and relevant management problems	How is the problem interpreted by the subjects involved? Is it grounded in the research case, is it generalizable?	Both the literature review and the evaluation interviews indicate that the problem is relevant in strategic management practice.
3.Design Evaluation	The utility, quality and efficacy of the artifact must be demonstrated through evaluation	What are the hermeneutic criteria needed to complete the evaluation?	The artifact is evaluated using both a quantitative and a qualitative method. They show the artifact's usefulness from both a structural and a user's point of view.
4.Research Contributions	Design research must provide clear and verifiable contributions	Are possible contradictions revealed between theoretical preconceptions and findings? Can the contribution be generalized?	The contribution is two-fold: on one hand, there is the collocational topic network that evolves during the iterations. On the other hand, the thesis contributes to general design theory by arguing that the dual nature of the artifact should be considered in the research process.
5.Research Rigor	Design research relies on rigorous methods in construction and evaluation of the artifact	What epistemological assumptions underlie the methods used? Are they inherently positivist?	The method developed is based on research in linguistics and assumes that contents of texts can be accurately visualized using statistical methods. It does, however, rely on user interpretations, in context, to be of use.
6.Design as a Search process	Design science is an iterative search process to discover a solution to a problem	(no additions to this guideline)	First, a basic version of the method is developed and evaluated. Then, with deeper understanding of the subject, an advanced version is developed and evaluated.
7.Communication of Research	Design research must be communicated to both technical and management focused audiences	To what extent does the communication pay attention do different interpretations by different addressees?	The research results have been communicated in IS journals and to those taking part in the evaluation interviews, These communications are quite different in nature.

Table 1. Design science guidelines by Hevner et al. (2004), interpretivist commentary by Niehaves (2007) and how they apply to this thesis

As Table 1 above shows, design science should not be considered a “third way”, an alternative epistemology to positivist or interpretivist epistemologies (as has been claimed by Vaishnavi & Kuehler 2004). Instead, it is usually conducted so that it carries the epistemological assumptions of either of these. Alternatively, if acknowledged by the design researcher, it could be conducted within some other epistemology, such as critical. Positivist, interpretivist and critical are the three main paradigms suggested for information systems by Orlikowski & Baroudi (1991).

This thesis also argues that design science is not an alternative to qualitative and quantitative research, although this has been suggested by Brian Smith in a panel paper (Purao et al. 2008). Instead, this thesis takes the standpoint that design science is a prescriptive research approach that can utilize both qualitative and quantitative research methods. In fact, the evaluation of the artifact developed in this thesis is carried out using both a qualitative and a quantitative method. This is discussed further in Chapter 5.

2.2.2 Design science and other prescriptive approaches to research

There are a number of other research approaches in management science that bear what Jönsson and Lukka (2007, 377) consider to be a Wittgensteinian family resemblance to design science: action research, constructive research, clinical research, and action science. All of these research approaches have as their goal to not just describe a phenomenon, but to change it, and are thus prescriptive rather than descriptive in nature (the term interventionist research is used by Jönsson and Lukka). This chapter contains a brief discussion on why this thesis falls under the category of design science, and should not be considered to represent any of the other prescriptive approaches.

Action research, with its origins in the writings of Kurt Lewin, has been widely utilized as a problem-solving paradigm in the behavioural sciences. Action

science also stems from this research tradition. Järvinen (2007) argues that the design science paradigm is similar to that of action research. Both approaches aim to contribute to practice as well as to theory. However, it can be argued that action research and design science research are not identical, as design science can be carried out in an artificial setting, not involving intervention, which is an essential part of action research (Purao et al.2008). For this reason, the research carried out in this thesis cannot be categorized as action research.

Constructive research, coined in Finnish management accounting research (Kasanen et al. 1993) has been particularly popular in Finland and also conducted to some extent in the other Nordic countries. The approach calls for strong intervention in the case company – the artifact is developed jointly with the case company. Clinical research, using an analogy from the medical sciences, also focuses on addressing and solving problems of a client organization. It puts less emphasis on theoretical issues and is strongly oriented towards the particulars of a case organization. In this thesis, a case company is used only for the qualitative evaluation. The artifact that is developed is intended for wider use and is not specific to the company in question.

As a conclusion, design science calls for less intervention than many other prescriptive research approaches. This means that the resulting artifact, in many cases, cannot be directly put into use in any single company without customization. However, the strength of the approach lies in the generalizability of the results.

2.3 The research process in design science

Hevner et al.(2004) outline a framework for the design science research process in information systems. The framework can be seen in Figure 2 below:

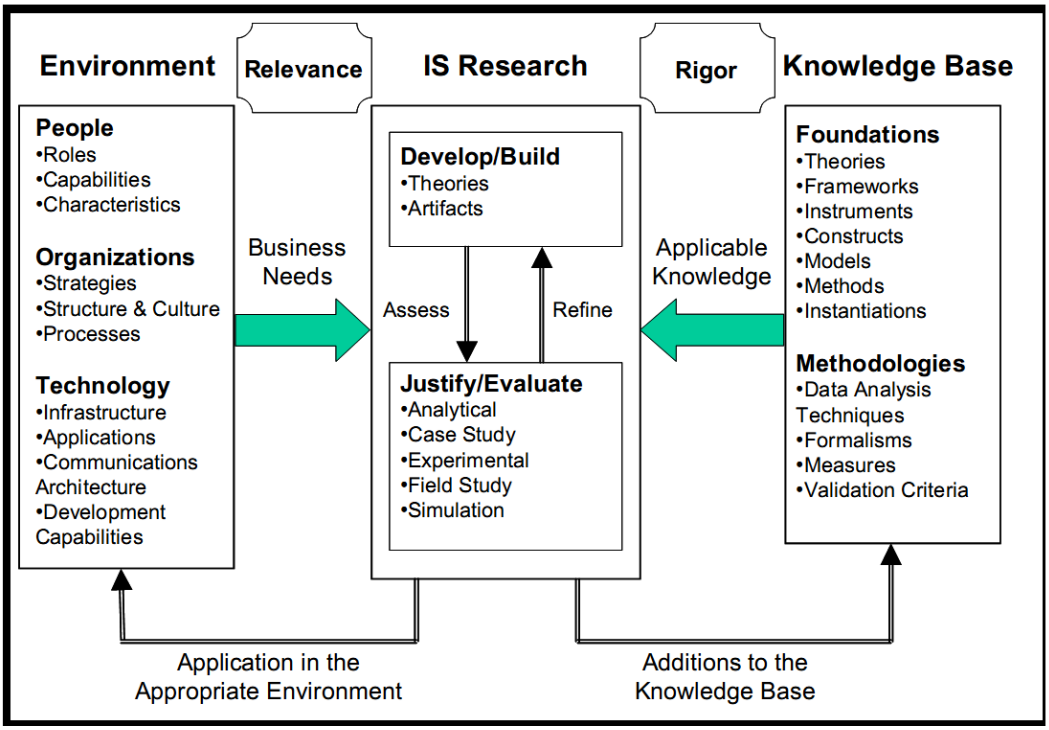

Figure 2. Information systems research framework (Hevner et al.2004)

The main aim of the process is to develop artifacts that are relevant to the needs of the *environment*, i.e. the business setting, encompassing people, organizations, and technology. These artifacts are created through the rigorous application of the knowledge in the *knowledge base*, i.e. the theoretical background of the artifact. The process, being iterative in nature, calls for repeated assessments of the utility of the artifact in its environment and improvements to it, accordingly.

2.3.1 Building an artifact

Once the business needs of the environment have been defined, the design science researcher must construct a solution that addresses the need and solves or in some way alleviates it. Simon (1996) coined the term *satisfice,* as a portmanteau word combining the words *satisfy* and *suffice*: a satisficing solution must be good enough to satisfy a need, but it is not necessarily the optimal solution. Simon takes into consideration the boundaries of human rationality and argues that optimal solutions cannot always be reached.

According to March and Smith (1995), there are four types of output in design science research: constructs, models, methods and instantiations. The artifact developed in this thesis is primarily a method, but methods are based on underlying constructs and models as well. These four types of output will be discussed in more detail next:

1. *Constructs*. These are concepts that form the vocabulary of a domain, such as a discipline or a sub-discipline. They are terms used to discuss problems and their solutions within a particular domain. This thesis introduces the vocabulary of collocational networks (concepts such as *node*, *topic word*) into the domain of competitive intelligence.
2. *Models*. These are sets of propositions that express relationships among constructs. The artifact developed in this thesis, the collocational topic network, is based on the idea of statistical co-occurrence of words as a model of the contents of a text. As such, it can be utilized for visualizing how certain topics are presented in texts.
3. *Methods*. These are sets of steps, such as algorithms or guidelines, used to perform a task. In this thesis, the collocational topic network is suggested as a text visualization method for competitive intelligence purposes, and thus falls primarily under this category.
4. *Instantiations*. These are realizations of the artifact in a real-world environment. Instantiations lie outside the scope of this thesis.

One further form of output in design research has been suggested by Purao 2002: *theories.* This thesis contributes to design theory through its argument that the dual nature of the artifact should be taken into account in design science activities. This is discussed in more depth in Chapter 2.4.

2.3.2 Evaluating the artifact

After a design artifact has been built, its utility, quality and efficacy must be rigorously demonstrated (Hevner et al.2004). This is done through evaluation. Hevner et al. (2004) suggest five types of evaluation methods: observational, analytical, experimental, testing methods, and descriptive methods. In this thesis, two evaluation methods are used. The first one, which consists of comparing changes in networks to changes in financial figures, can be described as an analytical method. The second method, qualitative interviews, falls under the category of observational methods. The evaluation is thus carried out both using a quantitative and a qualitative method, which complement each other.

The research papers that make up this thesis also reflect the iterative build/evaluate cycle of Hevner et al. (2004), as visualized in Figure 3 below:

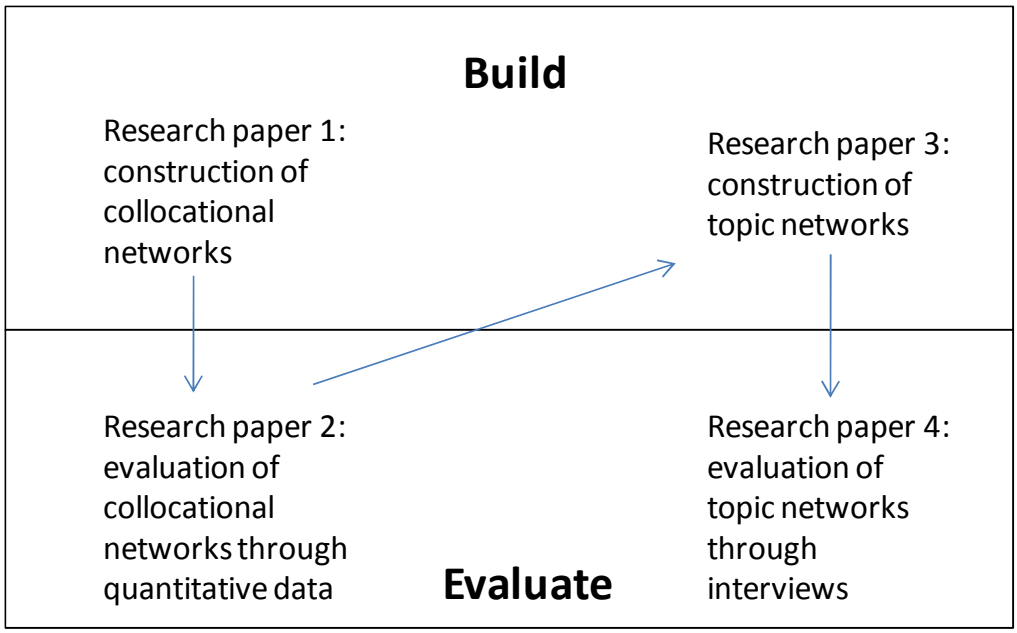

Figure 3. The research papers of this thesis reflecting the build/evaluate cycle by Hevner et al. (2004)

In the first research paper, collocational networks, a method for automatically visualizing changes in sequences of texts, is constructed. In the second paper, the method is evaluated in a quantitative manner. Changes in collocational networks produced out of quarterly report texts by stock-listed companies are compared to changes in self-organizing maps produced out of financial figures in the same reports.

The work on the second paper raised the question whether collocational networks on a specific user-defined topic could be even more useful in some cases. Thus, a process for creating such networks is constructed in the third research paper. Finally, the improved method is evaluated in the fourth paper using qualitative interviews with strategic decision makers in a case company.

2.4 Ontology of the design artifact

As Chapter 2.3 above reflects, much of the theoretical discussion within design science research in information systems has concentrated on the research process. There has been less theoretical discussion on the nature of artifacts produced in this process. There are, however, some exceptions within IS research: Orlikowski & Iacono (2001) argue that the theorizing of IT artifacts in context has been overlooked, as most research has focused on the context alone or on the technical capabilities of artifacts. This has lead to a simplified view of IT artifacts as "relatively stable, discrete, independent, and fixed" (Orlikowski and Iacono, 2001). Purao (2002) notes that because the artifact does not exist at the beginning of the design process, the researcher's ontological stance towards it will shift from emergent to realist as the artifact begins to take shape. At the same time, the environment for which the artifact is being designed assumes an emergent ontological stance, as it is moulded during the design process by the researcher. Herbert Simon's book The Sciences of the Artificial (1996), considered to be the work that design science originates from, also contains a view of the design

artifact that deserves to be discussed, particularly in the light of information systems. According to (Simon, 1996) an artifact:

> ...can be thought of as a meeting point – 'an interface' in today's terms – between an 'inner' environment, the substance and organization of the artifact itself, and an 'outer' environment, the surroundings in which it operates (p.6).

Simon thus emphasizes the role of the artifact as creating a relation between the natural and social worlds, in his words the inner and outer environments. Elaborating further on this, Kroes (2002) argues that technical artifacts can be considered to have a dual nature: on one hand, they are physical objects, on the other hand they are intentional objects, created to perform a certain function. This function is what separates artifacts from natural objects. Kroes approaches the issue from a design methodology point of view, but his arguments apply to design science research in information systems as well, as the specific nature of the physical artifact (for example, tangible object versus computer software) does not matter on this ontological level.

Kroes goes on to connect the idea of the technical artifact to three key notions, which are: the artifact's physical structure, its technical function and its context of intentional human action. The context of human action is brought into the argument, because a technical function would not exist outside a (human) intentional context. This is in line with the view expressed by Searle (1995) that functions are assigned to objects by conscious agents.

Kroes' argument differs from Simon's in that it is more explicit in articulating the existence of both a function and an intentional context. When Simon speaks of an artifact having a "goal", he seems to implicitly refer to a human context, but does not articulate it as such. The "outer environment" in Simon's argument does not correspond to a human context, but seems to encompass the world outside the artifact in general.

Kroes illustrates his argument through a figure similar to Figure 4 below. In Kroes' original figure, depicted as part A in the figure, the three dimensions of structure, function and context surround the artifact. Part B shows how Kroes' theory applies to the artifact that is constructed and evaluated in this thesis, i.e. collocational networks as a text visualization method for competitive intelligence purposes.

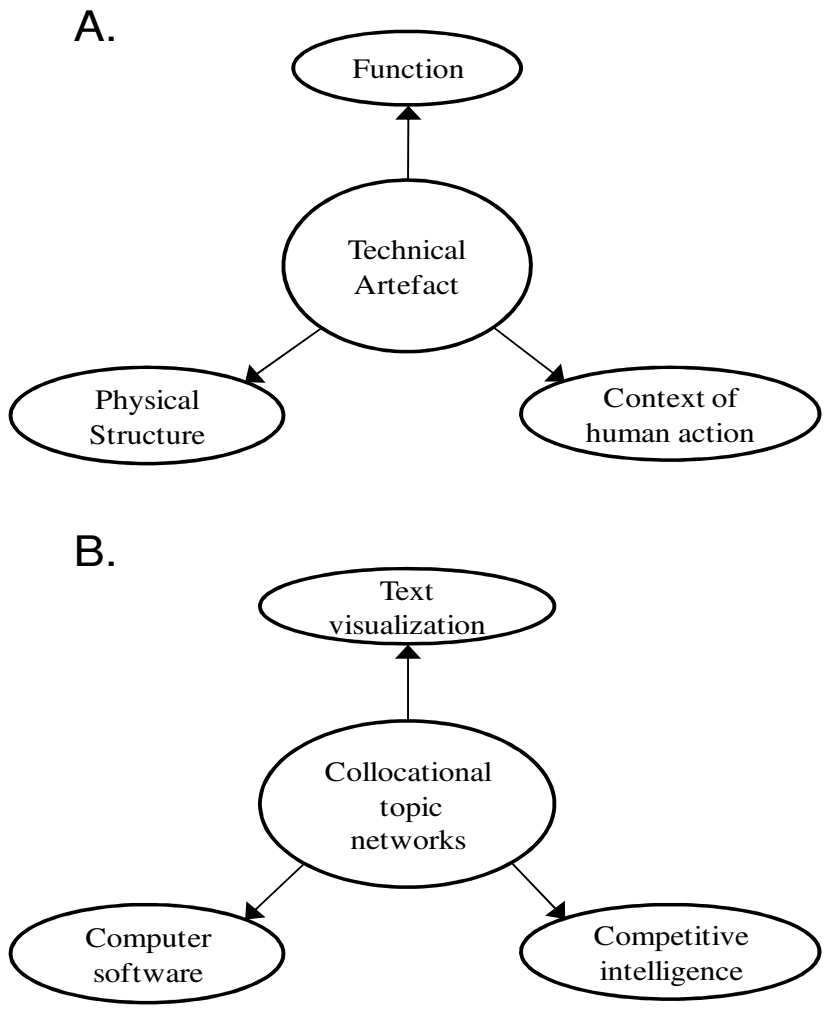

Figure 4. The nature of technical artifacts, exemplified by the artifact evaluated in this thesis. Based on Kroes (2002)

In this thesis, the development and evaluation of the artifact will attempt to cover the whole domain of its intentional aspect, i.e. it will take into account both the

function of the artifact (text visualization) and the context of human action in which it is applied (competitive intelligence). Both the artifact's context and its function will be discussed more thoroughly in Chapter 3. The principles behind the visualization method will be discussed in more detail in Chapter 4.

3 COMPETITIVE INTELLIGENCE AND STRATEGIC MANAGEMENT

3.1 Definitions of competitive intelligence

An often used definition of competitive intelligence (CI) is the one formulated by the Society of Competitive Intelligence Professionals (SCIP) which is as follows:

"a systematic and ethical program for gathering, analyzing, and managing any combination of data, information, and knowledge concerning the business environment in which a company operates that, when acted upon, will confer a significant competitive advantage or enable sound decisions to be made. Its primary role is strategic early warning."

The terminology within the field can be confusing, as many practitioners would probably rather use terms such as *business intelligence* or *market intelligence* for the definition presented above. Other terms used in connection with these kind of activities are *strategic intelligence* and *corporate intelligence.*
In this thesis, *competitive intelligence* is, following Fleisher (2003), used as a term covering both intelligence concerning competitors (the less often used *competitor intelligence* refers to this specific area) and intelligence concerning other players and markets (*market intelligence* often refers to this).

One important reason for making this choice is that within data and text mining terminology (see for example Sullivan 2001), the term *business intelligence* refers to intelligence derived from company internal data, whereas *competitive intelligence* refers to intelligence derived from data in the company's external

environment. In data mining, this distinction is of great importance, since a company is able to impose certain structures on internal data, at the time when it is being generated, whereas external data usually comes as completely unstructured documents. Also, the analysis of internal data usually serves operational purposes, whereas external data is typically used for forward-looking, strategic decision making. This difference between these two types of intelligence is of great significance from the point of view of analysis, yet it is rarely mentioned within competitive intelligence literature.

For the purposes of this thesis, then, the term *business intelligence* is reserved for analyzing company internal data and will not be extensively addressed in this study, although the text visualization methods presented here could in some cases also be applied to internal data. Note that the word *data* is used in this thesis to indicate raw data, i.e. letters and figures which as such carry no meaning, whereas the word *information* is reserved for data that carries meaning, given to it in a human context.

In this thesis, the possible users of the visualization method that is developed are named competitive intelligence practitioners, and the use of terms such as competitive intelligence managers or competitive intelligence analysts is avoided. This is done to reflect a current trend within corporations towards a system where different forms of competitive analysis are carried out independently by employees in several parts of the organization, not just by designated analysts within a competitive intelligence department.

How is intelligence work then conducted? Fleisher and Bensoussan (2007) present a model of the so-called intelligence cycle, which has five steps:

1 Planning the intelligence process
2 Collecting and processing data about the environment
3 Analyzing the collected data
4 Disseminating intelligence, i.e. presenting the insights to decision-makers

5 Evaluating the results through feedback from decision-makers

This model and a large part of today's competitive intelligence literature owes much to Porter's seminal work on competitive strategy (1980), which introduces the five forces that shape a company's competitive environment. The role of analysis in producing insights about these forces is crucial. The connection between competitive intelligence and strategic management will be discussed in more detail in the next chapter. Following that, the issue of overreliance on quantitative information in the analysis phase will be discussed in Chapter 3.3.

3.2 The role of competitive intelligence in strategic management

Competitive intelligence is needed in the operation of several functions within a company, for example marketing, sales, research, and HR. This thesis, however, focuses mainly on intelligence as a tool in strategic management. This is done partly because of space limitations, but also because the method created in this work is primarily suited for in-depth, forward-looking strategy work rather than for day-to-day competitive scanning.

Within the vast body of work that is strategic management literature, some theorists put more emphasis on knowing the external environment of a company than other theorists. This chapter will discuss some theories where this kind of knowledge, produced through competitive intelligence practice, as well as the use of qualitative information for this purpose, have a central role.

A useful way of illustrating differences between different theories of strategy is through the ten main schools of strategic management as defined by Mintzberg et al (1998). These can be seen in Table 2 below:

School	View of strategy process	Nature of school	Examples of well-known theorists	Role of the external environment and, implicitly, the focus of competitive intelligence
Design school	A process of conception	Prescriptive	Selznick	The threats and opportunities posed by the environment should be known
Planning school	A formal process	Prescriptive	Ansoff	A forward-looking audit to assess the external conditions of the organization is crucial
Positioning school	An analytical process	Prescriptive	Porter	A competitive analysis is crucial in order to find a position in the industry
Entrepreneurial school	A visionary process	Descriptive	Schumpeter	An entrepreneurial organization constantly needs to look for new opportunities
Cognitive school	A mental process	Descriptive	Simon	The environment is a construct produced by managerial beliefs
Learning school	An emergent process	Descriptive	Nonaka & Takeuchi	An organization needs constantly to be looking for knowledge outside its boundaries
Power school	A process of negotiation	Descriptive	Pfeffer & Salancik	An organization can act upon and negotiate with its environment
Cultural school	A collective process	Descriptive	Normann	The organization sees the environment through its own culture
Environmental school	A reactive process	Descriptive	Hannan & Freeman	The organization needs to know its environment in order to adapt to it
Configuration school	A process of transformation	-	Mintzberg	The organization's relation to its environment is defined by the stage the organization is in and to what stage it is moving.

Table 2. Ten schools of strategic management, adapted from Mintzberg et al. (1998)

The first three of these schools are what Mintzberg et al (1998) call prescriptive in nature. They aim to give directions as to how strategies should be formulated, whereas the next six schools are descriptive in nature, and focus on describing how strategies are made. The final school, called configuration, combines features of the others to examine different stages in company strategy and the transformation from one stage to another. Although most of these schools do in some way imply the importance of knowing the company-external world and thus using competitive intelligence (see the rightmost column in Table 2), only some of them address the importance of qualitative information for this purpose. Next, some of the most influential representatives of such schools of strategic management will be discussed.

Within competitive intelligence, the best-known strategist is likely to be Porter (1980), whom Mintzberg et al. (1998) place in the *positioning school*. The main argument of the positioning school can be summarized using a quote from Porter (1980:3):"the essence of formulating competitive strategy is relating a company to its environment". Porter argues that a company needs to find a position within its industry where it can best defend itself against external forces, or influence them. Particularly important external forces are the five competitive forces inside the industry in which a company operates. These are:

1 Rivalry among existing competitors in the market
2 Threat of new entrants to the market
3 Threat of substitute products
4 Bargaining power of suppliers
5 Bargaining power of buyers

Particularly for competitors and new entrants, Porter suggests that an analysis is carried out consisting of the following diagnostic components: the competitors' future goals, their current strategy, their assumptions about themselves and the industry, and their capabilities (encompassing both strengths and weaknesses). Porter remarks that most companies usually develop a sense for their competitors'

current strategies as well as their strengths and weaknesses, but give less attention to their future goals and the assumptions that drive their behaviour. From the point of view of this thesis this remark is important, because future goals and assumptions about the industry are the type of information that listed companies communicate to their shareholders in annual and quarterly reports. Exactly this kind of public qualitative information is visualized for competitive intelligence purposes in this thesis.

More recent representatives of the prescriptive positioning school include Kim & Mauborgne (2005), whose Blue Ocean strategy has received much interest among practitioners of strategic management in the 2000s. They argue that a company should give up the pursuit for market share in a metaphorical ocean of fierce competition turned blood red, and should instead focus on finding a niche, a blue ocean, where competition does not exist. For a company to be able to find such a blue ocean, a thorough analysis of the strategies of competitors is needed. Kim & Mauborgne (2005) present a number of tools for competitive analysis, one of them being the strategy canvas. This is a tool into which companies enter key factors of competition in their industry and then assess to what extent their competitors realize each factor. The aim is to find a combination of factors that does not yet exist within the industry. Although this kind of competitor analysis in its basic form focuses on the competitors' existing strategies, their plans for the future and the assumptions that drive them need to be taken into account as well, as competitors may be looking for that blue ocean too. Again, this is something that only qualitative information can provide insight into.

A key argument made by Nonaka & Takeuchi (1995), representing the *learning school*, is the distinction between tacit and explicit knowledge. Whereas tacit knowledge, according to their definition, is "personal, context-specific and therefore hard to formalize and communicate", explicit knowledge is "transmittable in formal, systematic language". One of the most important knowledge-creation processes in Nonaka & Takeuchi's theory is that of turning

tacit knowledge into explicit, through a process called externalization. A text visualization method, such as the one presented in this thesis can help in the externalization process by making changes and trends in the environment explicit and discussable within the organization and thus something that can be acted upon.

Although the strategic management theorists discussed above see the importance of including qualitative information in analysis of the organization's external environment, in practice, this is often not the case. Many companies still rely too extensively on quantitative information, while ignoring qualitative information. This issue will be discussed next.

3.3 Over-emphasis on quantitative information in intelligence and strategic management

Within both competitive intelligence and strategic management research, concern has been expressed for the quality of analysis in current intelligence practice. There are some issues that make this process very demanding for companies in today's world.

First, there is information overload. Companies are overwhelmed by the amount of information that is publicly available about their competitors, customers and suppliers. They need tools to be able to distinguish the relevant from the irrelevant.

Second, competitive analysis often relies heavily on quantitative financial information. Fleisher and Bensoussan (2003) call this phenomenon "ratio blinders". According to them, many organizations make the mistake of relying too much on financial information of their business environment. This can lead to a situation where companies can see a financial gap between their organization and a competitor but cannot see the reasons behind it and thus have not got the means

to close it. This issue relates to a mistake that is commonly made, the confusion of operational information with strategic information, mentioned by Zahra and Chaples (1993). Mintzberg (1994) argues that one of the fundamental fallacies of strategic planning is trusting exclusively on what he calls "hard" information. Although the term *hard* cannot be directly equated with *quantitative*, the connection is obvious: all quantitative information is not hard information, but all hard information has been quantified in some way. This leads to a limited view of the business environment: "a bias toward the quantitative can allow the economic to displace the social and the financial to displace the creative" (Mintzberg 1994, 258). According to Mintzberg, the main limitations of hard information are:

1 hard information is often limited in scope,
2 much hard information is too aggregated for effective use in strategy
3 much hard information arrives too late to be of use in strategy making
4 a surprising amount of hard information is unreliable

Mintzberg admits that soft information has it problems too, as it can be speculative, rely too much on human memory and be subject to what Mintzberg calls "all kinds of psychological distortions". Nevertheless, Mintzberg believes that strategic management should ideally draw on both hard and soft information. He sees hard information as informing the intellect, but soft information as generating wisdom.

Related to this is the view expressed by Pihlanto (2009), who, coming from a background in management accounting but discussing decision making in general, argues for a holistic approach to the decision maker. Such a view sees the decision maker primarily as a human individual, operating within three basic dimensions: consciousness, situationality, and corporeality. From a competitive intelligence point of view, consciousness and situationality are particularly important. Human consciousness is seen to produce understanding of objects and phenomena in a particular decision making situation. Because understanding presupposes that meaning has been given to these objects and phenomena, it can be inferred that qualitative information, e.g. on the causes or implications of a phenomenon, is

needed to produce true understanding of any phenomenon within the business environment. The situationality of the individual means that each decision making situation is unique, and new information is interpreted in relation to earlier situations – thus creating significant differences in interpretation between individuals. This approach leads to the view that quantitative information cannot be considered to be a fundamentally more exact and unambiguous base for decision making than qualitative information, as the interpretation of it is also always connected to a specific situation.

In order to complement and deepen intelligence based on quantitative financial information there seems to be a demand for methods that allow decision makers to systematically include qualitative company-external information in their analysis. Such information is available e.g. in publicly available texts produced by companies such as competitors, suppliers or partners. Most competitive intelligence practitioners today do read such texts, but in order to incorporate the information effectively into the analysis process, a systematic methodology for analyzing large quantities of text is needed. Working in complex business environments, with a high number of companies that need to be analyzed and an ever-increasing amount of texts being produced, a manual scan quickly becomes difficult.

Text visualization, however, promises to alleviate this task. Recent studies on the effects of information visualization in a management context (Eppler and Platts 2009, Lurie and Mason 2007, Eppler 2006) indicate that visualizations can greatly assist decision makers by providing new insights, increasing the accessibility of information and facilitating the synthesis of information from different sources. For the visualization of qualitative information in strategic management, Eppler and Platts (2009) suggest traditional methods such as the Boston Consulting Group matrix or the SWOT matrix. This thesis argues that automatic methods for text visualization, such as the one presented in this thesis, should also be used, particularly for large text masses.

In the next chapter, text visualization and its background in text mining will be discussed, and collocational networks will be suggested as a method for visualizing changes in sequences of texts produced in the external environment of a company.

4 COLLOCATIONAL NETWORKS AS A TEXT VISUALIZATION TOOL

4.1 Background on text mining and visualization

Text mining, or text data mining, can be defined as the discovery of trends and patterns within textual data (Hearst, 1999). *Text visualization*, also known as visual text mining, aims at producing visual results of a text mining process. This typically makes the end results of the process more accessible and user-friendly.
The background of the text mining method presented in this thesis lies in corpus linguistics, a sub-discipline of linguistics that deals with the statistical analysis of large collections of texts, also known as corpora. The main aim of corpus linguistics is to discover patterns of language usage with the help of pre-defined general text collections containing millions of words or collections of genre-specific texts (Biber et al, 1998; Sinclair, 1991). Corpus linguistic methods have been applied to various linguistic activities, including language teaching, the production of dictionaries and grammars, literary studies and translation (Hunston, 2002).

In recent years, some attempts have been made to introduce text mining methods to deal with competitive intelligence issues. For example, Leong et al. (2004) use a text mining approach to analyse the textual content of several university Web sites to uncover the persuasive themes that these organisations use to attract prospective students. Fattori et al. (2003) compare a text mining tool with traditional portfolio analysis methods within the context of patent analysis. Courseault Trumbach et al. (2006) present a method that allows small technology

companies with limited resources for competitive intelligence to keep abreast with scientific and technological developments in their field using text mining.

There are currently a number of commercial text mining software packages available which also carry visualization features. There are also a number of non-commercial text visualization tools online that can be freely used. A non-comprehensive list of the main alternatives to the method presented in this thesis, both commercial and non-commercial, can be seen in table 3 below. The list has been compiled through online searches for “text visualization” and through recent research papers on text visualization tools (Yang et al. 2008, Bose 2008). Tools mainly aimed at visualizing web content have been omitted from the list, as the focus in this study is on visualizing a series of individual static documents. The data in table 3 is based on the vendors’ presentations of the products and use of the tools where possible.

Alternative text visualization tools				
Name of product	**Website**	**License type**	**Visualization features**	**Notes**
Temis Luxid for Competitive Intelligence	www.temis.com	Commercial	Visual clustering of words	
SAS Text Miner	www.sas.com	Commercial	Concept Linking: displays words in clusters	
OmniViz	www.biowisdom.com	Commercial	Visual word clustering from collections of text documents	Mainly aimed at healthcare industry
SWAPit DocMINER	www.fit.fraunhofer.de	Commercial	Creates document maps from collections of text documents	
IBM ManyEyes	manyeyes.alphaworks.ibm.com	Free to use on site	Tag Cloud: emphasizes frequent words in a text Word Tree: shows text as trees, based on a search word Phrase Net: produces a network of words appearing in pre-defined phrases Word Cloud Generator: shows frequent words in a text as a cloud	
TAPoRware	taporware.ualberta.ca	Free to use on site, download available for educational use	Raining Words: Words move differently based on their frequency Visual Collocator: produces visualizations of collocations Weighted Centroid: a circular graph base on word frequency	
Wordle	www.wordle.net	Free to use on site	Word cloud: shows frequent words in a text as a cloud	Focus is on visual effect

Table 3. List of some alternative text visualization tools

Many of the commercial products, in particular, have features that fall under the category of clustering (Bose 2008) i.e. they group documents according to their content. This is something that was not attempted in the current study, but which could be considered useful from a competitive intelligence point of view: clustering companies according to topics in their communications material.

However, the aim of this thesis is not to present a text visualization method to compete with commercial products, but to encourage the use of textual material in competitive intelligence through the use of a simple tool. The tools should provide an easily interpretable output in order to be useful for users without a background in linguistics or statistics. Although commercial text mining and text visualization products have been available for a long time, they are often considered by

potential users to be difficult to grasp, “black box methods”, as noted by Krier and Zaccá (2002). This is one reason why such packages are rarely found in companies today. As an example, none of the interviewees in the evaluation interview reported in Chapter 6 had encountered such tools during their working life.

Some of the free tools, on the other hand, are widely used online. Word clouds, for example, are now being used on numerous web sites as a tool for search. The challenge lies in seeing these tools as not just toys (as wordle.com calls itself), but as aids for discovering interesting patterns in text. This study takes what could be a free tool, collocational networks, and aims to introduce it for corporate use through evaluation and further development within the context of competitive intelligence. The networks are visualizations of central concepts in a text. Based on its description, the TAPoRWare Visual Collocator sounds like a similar tool, but it did not function at the time the site was visited.

Collocational networks work particularly well for visualizing changes in sequences of texts of similar structure, and can easily be interpreted by users with no previous knowledge of linguistics. The principles behind this method will be discussed in the next chapter.

4.2 Linguistic origins of collocational networks

For competitive intelligence practitioners reading texts produced in their business environment, trends and changes around specific topics are particularly important. Automatic discovery of the topics or subject matters of a text through linguistic methods has been explored extensively by Phillips (1985). He discusses three approaches based in linguistic theory, which are relevant components in creating a methodology for analysing subject matter in a text. Firstly, there is the classic Saussurean approach, which emphasizes the arbitrariness of the linguistic sign

(Saussure, 1995, originally published 1916). As the relationship between the signifier (i.e., the word) and signified (i.e. the real-world object that a word refers to) is arbitrary, this approach suggests that there cannot be a direct correspondence between the systems of language and real-world phenomena. This way of thinking is not directly useful for the development of new methods, but rather reminds the creator of any text mining method of its limitations. A text can never be expected to reflect reality unambiguously.

Secondly, Phillips mentions the Firthian approach, which emphasizes meaning as a function in context (Firth, 1957). The concept "context of situation" is separated from textual context, which is directly susceptible to linguistic analysis. The textual context is considered to be actualised through lexical patterning such as collocation, the co-occurrence of words. Collocation will be discussed in more detail in Chapter 4.2.1.

Thirdly, there is the so called text linguistic approach, which distinguishes between two levels of meaning in discourse. These levels are on one hand the level of sentences or sequences of sentences and on the other hand the level of parts of the discourse or the discourse as a whole. This view suggests that the researcher looking at patterning in a text should not forget the significance of the text as a whole and its role in a particular situation.

All these three approaches present some conceptions about the relationship between a text and the non-linguistic world. Using these approaches as a starting point, Phillips suggests what he calls a knowledge-free analysis (not using other resources than the text) of the words in a text. An analysis of this kind should reveal systematic textual patterning, which in turn contributes to the semantic structure of the text and functions as a basis for the emergence of subject matter.

Williams (1998), whose article is the main source for the method presented in this thesis, collocational networks, draws some of his ideas from Phillips' study. His

work is slightly different, though, as his aim is to find words central to a particular sublanguage instead of words central to a particular text. This leads him to produce networks out of text corpora rather than individual texts. This thesis constitutes a return to Phillips' original pursuit of the subject matter of a text. The method allows for creation of networks consisting of words that occur together, in a statistically significant way, in an individual text or in a collection of texts. The details of calculating such co-occurrence, also known as collocation, will be discussed next.

4.2.1 Definitions of collocation

In this thesis, collocation is primarily interpreted simply as "the occurrence of two or more words within a short space of each other in a text" following Sinclair (1991,170), a central work within corpus linguistics. As the results of this study will show, collocations produced by the method used here will often also converge with the more specific definition presented by Kjellmer (1987,133): "a sequence of words that occurs more than once in identical form". What makes collocation interesting is the fact that it provides an opportunity for studying the meaning of a word by exploring the environment in which the word occurs. The meaning of a single word studied in isolation in a particular text is difficult to define, as meaning is cumulated through the co-occurrence of words.

It should be noted that no attempt has been made to draw a line between collocations and compound words. For this reason some pairs of words which in another context would be called compounds will be treated as collocations in this study. Usually collocation is treated as the co-occurrence of two words, which form a collocational pair. The approach used in this study gives us the possibility of examining larger frameworks by creating networks of collocates, instead of just concentrating on pairs.

It is important to note the definition of *word* used in this thesis. Following Williams (1998) no lemmatisation (i.e. grouping inflected versions of a word-

form into one word) has been carried out, as different inflected or derived forms might lead to differences in the frame of reference. Therefore each orthographic word-form represents a separate word.

4.2.2 Measuring significant collocation

A central factor in the collocational networks method is the concept of significant collocation. Significant collocation takes place when two or more words occur together more frequently than would be expected by co-incidence. There are several ways of measuring the significance of collocation, but for the purposes of this thesis, the measure that is mainly used is the *Mutual Information* (*MI) score.* This choice is discussed further below.

The MI score is a concept originating in information theory, and it was proposed for linguistic purposes by Church and Hanks (1990). It compares the frequency of co-occurrence of two words, the node word and its collocate, with the frequency of their occurrence independently of each other.

The MI score is calculated as follows:

$$MI(n,c) = \log_2 \frac{f(n,c) \times N}{f(n)f(c)}$$

where n stands for node, c for collocate and N for the size of the text in number of words. The higher the MI score, the more significant the collocation between these two words is. By selecting a word in a text and calculating the words that produce the highest MI-scores with it, the beginnings of a collocational network are made. Figure 5 below shows an example of such a network.

Corpus linguistics literature knows a number of different measures for such co-occurrences of words. For example, Pecina (2008) lists and compares 55 different

measures for purposes of multiword expression extraction in German and Czech texts. His recommendation is to combine several measures for best results. However, the purpose of using co-occurrence measures in this thesis is quite different: the aim is not to find word pairs that native speakers would consider multiword expressions (such as compound words), but to find pairs of words that occur closely together in a particular text, giving an indication of the context in which a node word is used. The evaluation of different measures is therefore much more difficult, as there is no absolute definition of what should be included in the context. Mutual information has mainly been used as a measure in this thesis because of its dominant position in contemporary corpus linguistics and its proven usefulness in a variety of cases (for examples, see Oakes 1998). The use of MI does, however, have drawbacks. Stubbs (1995) argues that the MI score has to be interpreted with care. It is non-directional, and therefore has the same value regardless of which word of a pair is the collocate and which is the node. The MI score is also sensitive to changes in the absolute number of collocates, when the relative proportion of joint occurrences compared to independent occurrences remains the same. In these cases it works "counter-intuitively" (Stubbs 1995), decreasing as the absolute number of collocates increases. This means that two words that always occur together in a text get a higher MI score if they occur only once than if they occur more frequently.

Another commonly used measure for co-occurrence, the Z-score (Berry-Rogghe 1973), was used in the third research paper (Magnusson 2010), as it was included in the TAPoRware tools. Based on this study, no significant practical differences were noted between the Z-score and the MI score. When developing the method further, the selection of co-occurrence measure is certainly an issue to be explored more thoroughly.

4.3 Producing a collocational network

The network in Figure 5, based on the text of Nokia's first quarter report in 2001, has been created around the central node *Nokia*, which is shown here connected to eight of its most significant collocates: *networks*, *mobile*, *ventures* etc. Some of these collocates are linked to a number of collocates of their own. The node word in basic collocational networks is the word with the highest frequency in the text. The frequency of each word is indicated in brackets after the word and the MI score is marked on the line connecting the words. This particular network will be discussed in more detail in Chapter 5.2.

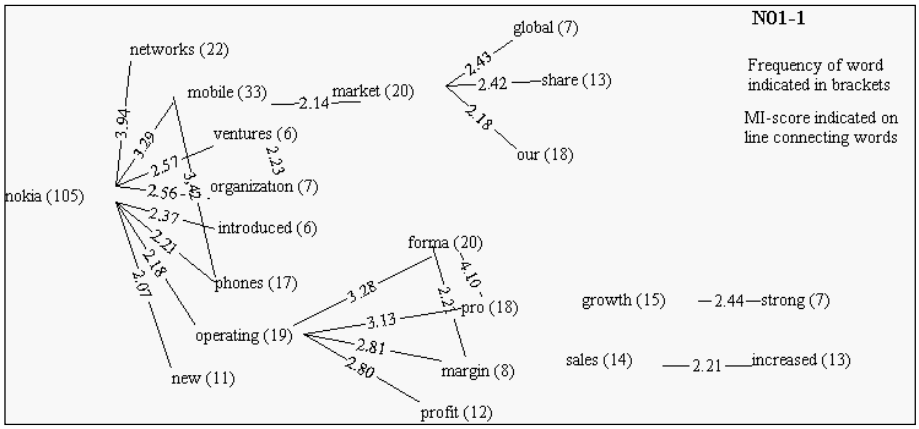

Figure 5. An example of a collocational network (Magnusson et al. 2005)

Figure 5 exemplifies the basic structure of a collocational network: it is a two-dimensional representation of the node word in a text, surrounded by its main collocates, i.e. the words that appear in a statistically significant way together with the node word in the text. The node word is the word with the highest frequency in the text (in the case of basic collocational networks) or any word in the text of particular interest to the user (in the case of collocational topic networks). The method allows for drawing as many or as few collocates as is relevant in each case. The collocate words are then surrounded by their most statistically significant collocates.

During the research process a number of tools for assisting in the creation of networks were tried. At the very beginning, collocations with the highest MI score

were calculated using an AWK script in a Unix environment, the result of which was then fed into Graphviz Neato, an open source software product (available at http://www.graphviz.org), for the automatic production of networks. For more ease of use, the process was then moved to a Microsoft Windows environment, where the calculation of MI scores was done using the freeware software AntConc (available at http://www.antlab.sci.waseda.ac.jp/antconc_index.html). The networks were then manually drawn using Microsoft PowerPoint. The free TAPoRware online tools provided by McMaster University, were also tried, but the performance of the site (http://taporware.mcmaster.ca/) was found to be unreliable and attempts to contact site maintenance failed. This service has later been moved to the University of Alberta (http://taporware.ualberta.ca). Although the production of networks was slowed down by manual drawing, it gave the user more freedom in designing the look of the networks. The use of a collection of tools was not considered a problem, since the research process was exploratory, aiming to evaluate whether visualizations of this kind would be useful for competitive intelligence practitioners, not to produce a finished product for commercial or even scholarly use at this stage.

4.4 Collocational networks as a visualization tool for sequences of texts

Collocational networks are particularly suitable for temporal text mining. This is an umbrella term for methods used to discover temporal patterns in texts collected over time (Mei and Zhai, 2005). For the purposes of competitive intelligence, temporal text mining looks promising, because an important part of competitive intelligence analysis lies in understanding not only competitor's current strategic decisions but their development over time. This also creates a possibility of more accurately predicting their future decisions.

Collocational networks show which words occur, in a statistically significant way, together with a particular node word in a text. This thesis assumes that these co-occurrences reflect the context in which that particular word is presented in the text. Thus, when analyzing a sequence of texts, changes in occurring words from one network to the next will reflect changes in the context in which the node word occurs in the original texts. This thesis focuses on corporate annual reports and quarterly reports as material for temporal text visualization, but other types of text can also be visualized with this method.

4.4.1 Corporate financial report texts as material for visualization

In this thesis, quarterly and annual reports from public companies are used as research material. Such reports are ideal material for temporal competitive analysis, as they reflect changes in a company's strategy and competitive situation over time. From a text mining point of view, they are also suitable for the task, as there is usually a similar document structure (similar passages, similar length) from one quarter to the next. For examples of collocational networks created out of financial reports over a period of six years, see Figure 13 in Chapter 6. The figure shows how changes take place around the topic of *service*, with new words appearing and others disappearing each year.

However, it should be kept in mind that financial report texts, although required by law to be truthful, are intended to produce a positive view of the company in the mind of shareholders and other groups of stakeholders. Michalisin and White (2000) give some reasons for why the contents of an annual report text can never be considered entirely factual: there are less strict reporting standards for annual report text, whereas the figures presented in annual reports have to be accurate. Another reason is that managers have an incentive to write text that emphasizes things that are positive from the company's point of view. Even when managers make a conscious effort not to do this, they still lack the objectivity that is needed in order for the text to be an accurate description of the company. And finally,

public relations professionals are hired by some companies to write annual report texts - a clear indication that corporate reports are used for PR purposes. This is why the text in any corporate report should be seen as a means of communication rather than as a piece of exact information about a company.

It is a common belief that companies, when given the chance, always emphasize positive events and play down anything that might be seen as negative, but as Crombie and Samujh (1999) point out, even bad news can be used strategically. Their study concentrated on a single letter to the shareholders in an annual report of a small New Zealand company. In this letter the chairman addresses a number of problems that the company has faced during the past year. A closer study of the company reveals that these problems are minor ones, used by the chairman to distract the attention of the readers from more serious issues while simultaneously presenting the management positively as problem solvers.

Some studies show that it is possible to find a direct connection between financial report text and the financial performance of the company. Osborne et al. (2001), working within the field of management studies, assume that the text in the chairman's letter to the shareholders, included in most financial reports, reflects the strategic thinking of the management. In their study different companies were clustered according to performance on one hand and according to themes in the chairman's letters on the other, and these clusters were seen to converge in the study.

In competitive intelligence practice today, corporate reports do seem to play an important part: not only do representatives of companies read them to obtain more information about their competitors, many companies also carefully note the kind of disclosures their competitors make (Meek et al. 1995). Further support for the use of corporate report texts as a source for competitive intelligence comes from the observation made by Clark Williams (2008) that corporate reporting is changing: companies have begun to use voluntary disclosures more often, discussing matters that are not required by law to be included in their reports.

There is also an increasing tendency to combine various types of disclosure (such as social and environmental information) and to present them in the context of financial data. Nevertheless, voluntary disclosures are typically used in order to present the company in a positive light. This means that corporate reports (and visualizations of them) should always be interpreted with the communicative aims of the organization that produced them in mind.

4.5 Developing the method further: Collocational topic networks

The collocational networks produced for the first evaluation (discussed in more detail in Chapter 5), have the most frequent word of each text as their central node. As this method produces networks that centre around something that is frequently mentioned in the text, it gives the user an overall picture of the main topic of the text (for example, see the network in Figure 5). After this type of method had been developed and evaluated, it was further developed into another form of visualization called *collocational topic networks*, or *topic networks* for short. Technically, topic networks do not differ from basic collocational networks, as the algorithm used and the construction method is the same. However, for such networks, the selection of the central node is done by users and is thus not connected to word frequency. This gives the users a greater freedom in exploring topics that may be of interest to them, but which are not represented by high-frequency words in the text.

The topic for a topic network could be selected by a competitive intelligence practitioner or through an expert elicitation process, as was done in the third research paper of this thesis (Magnusson 2010). In the study, stock market analysts were asked to read financial reports and point out subjects that they found to be of greatest importance. Matters relating to competition were the ones that

interested most respondents in that study, and thus the word *competition* was selected to be the central node of the topic networks produced.

For the purposes of the second evaluation, discussed in more detail in Chapter 6, collocational topic networks were created out of the word *service*, as this was a topic that was expected to concern all interviewees taking part in the evaluation.

4.6 Interpreting collocational networks

As discussed earlier, collocational networks show the words that co-occur with the node word in a text. The co-occurrences reflect the context in which the node word is presented in the text. The key to interpretation therefore lies in discovering changes in sequences of networks. What new words appear in the context of the node word? What words disappear? The method allows for a number of usage scenarios, depending on the strategic decision making tasks that the networks are being used for. As an example, the networks could be analyzed for the tone of appearing and disappearing words. Is it negative or positive? For the case companies discussed in more detail in Chapter 5, words indicating revenue growth begin to disappear from the networks, to be replaced by words indicating a decline in sales and even financial difficulties. As that particular case shows, this is an indication of changes in the financial status of the companies. To assist with spotting such patterns, a table such as Table 4 can be produced, showing exactly in which network a word appears or disappears.

A slightly different approach is taken in the case presented in Chapter 6. There, changes in the context of the word “service” are observed. The visualized text material is a combination of reports from several companies, so these changes can be seen to reflect the evolution of the service concept within an industry.

The networks can be drawn so that they indicate both word frequencies and the strength of the collocation as expressed by the MI score. It should be noted that

the inclusion of such figures can give the users a false impression of exactness, and this is something to be avoided when working with qualitative material. Language use and the interpretation of meaning can never be exact. However, in order to produce the networks, or any other form of text visualization, some degree of quantification is by definition imposed on the text, and meaningful content, available in the original text, is lost in the process. This is why the networks should be used as a starting point for discussion and analysis, not be seen as the result of analysis, a mistake that is often made when working with text mining and visualization tools (Bose 2008).

The material used for visualization obviously affects the interpretation. As discussed in Chapter 4.4.1, corporate report texts are used for promoting the company's views and its public image, and should thus not be considered objective accounts of what is happening in the company. This applies to visualizations of the texts as well. A company may choose to publicly emphasize a certain aspect of a topic, which is then likely to turn up in the networks, whereas another aspect, just as relevant for the company, is never mentioned, and thus never turns up in the networks. When visualizing material produced by companies themselves, the aim of the analysis should be to discover the context in which companies are discussing some particular topic (or the industry is discussing, as is the case in the evaluation case discussed in Chapter 6). This information can then be used for making strategic decisions on how to approach these companies (if they are customers or partners) or how to stand out from them in the eyes of customers (if they are competitors).

5 QUANTITATIVE EVALUATION OF COLLOCATIONAL NETWORKS

Evaluation plays a crucial role in design science research. In this thesis, the artifact is evaluated twice, using two different evaluation methods. Between the evaluations, the artifact was developed to fit user needs better, so that it visualizes textual changes concerning a certain topic (as described in Chapter 4.4), instead of visualizing changes in the main topic of the document. First, a quantitative evaluation of the method is carried out. The results of the evaluation are presented in this chapter. The second evaluation, which is of a qualitative nature, is presented in Chapter 6.

5.1. Background and data selection

In order to evaluate whether the collocational networks reflect the underlying texts, and whether they thus can be useful for forward-looking competitive intelligence purposes, an evaluation setting containing a comparison of collocational networks with quantitative financial data was constructed. For this evaluation, the textual data and the financial data from the quarterly reports of three mobile handset vendors was visualized. The following hypothesis, in line with Osborne et al (2001), was made:

When a company's external environment and internal performance potential change:

- there may be a change in the strategic thinking, expectations, and associated planning by its management,

- this will then be reflected immediately in company texts, for example in its quarterly reports, and
- this will also be reflected with some time lag in the financial performance data; i.e. it takes time for the expectations and actions of the corporate management to materialize in the financial results.

The time lag was expected because of the different nature of the quantitative and qualitative data in the quarterly reports. The financial performance figures can only represent developments prior to the publication of the quarterly report, whereas predictions and forward-looking assessments can be stated in the text.

The data used for this study consisted of the quarterly reports of mobile handset vendors Nokia Corporation (Nokia), Motorola, Inc. (Motorola), and Telefonaktiebolaget LM Ericsson (Ericsson) over the years 2000-2001. This choice was made because the years 2000 and 2001 saw a dramatic downturn in the growth of the IT industry. This period was therefore considered as most suitable for studying changes, as they would be found in the financial figures of most IT companies at that time. The telecommunications industry in particular had experienced very strong growth in the 1990s.

Collocational networks were produced out of the texts of each quarterly report produced by the three companies during the period under study. First, the MI score for all words occurring within a span of four from one another was calculated. With text sizes of approximately 4,000 words, a minimum MI score of 2.00 was found to produce a network of a suitable size. Lowering the score would have brought in words that occur together only occasionally, while a higher limit would have produced a network with only the most frequent combinations, leaving out interesting changes among the mid-frequency words. Words with little relevance (prepositions, articles, conjunctions, words referring to the time span of the report, and figures or currency) were left out.

In order to visualize the changes in the financial ratios of the quarterly reports, a self-organizing map (SOM) was created. A SOM is an unsupervised neural network that maps multidimensional data onto a two-dimensional topological map; it is commonly used for exploratory data mining. In this study, it was applied to corporate benchmarking, where it has been found by managers to be a feasible tool for strategic decision making, as shown by Eklund (2004). The strength of the SOM lies in its capability to visualize multiple financial ratios simultaneously.

The SOM clusters data according to similarities, displaying the result as a map of nodes separated by borders: dark ones represent great differences, while light ones indicate similarities, forming clusters of data. In this study a SOM was used to benchmark the performance of 88 international telecommunications companies. A number of financial ratios were calculated, after which the SOM map was trained. The ratios included were: operating margin, return on equity, return on total assets (profitability ratios), current ratio (liquidity ratio), equity to capital, interest coverage (solvency ratios), and receivables turnover (efficiency ratio). These ratios were chosen based on an empirical study on the reliability and validity of financial ratios in international comparisons (Lehtinen 1996). The map created was then used to visualize the actual financial performance of the three companies studied, so that the performance could be compared to the collocational networks.

Further details of self-organizing maps lie beyond the scope of this thesis, as the method was not used for text visualization, but a thorough description of SOMs and their application to corporate benchmarking can be found in Eklund (2004).

5.2. Analysis of the material

In the collocational networks made out of Nokia's quarterly reports, a long period of stability was seen. During the year 2000, the networks hardly changed. They were almost identical, containing the name *Nokia* as a central node with links to

other nodes referring to the company's business segments, such as *Networks* or *Mobile Phones*, or general nouns used in business, such as *sales*, *market*, and *growth*. However, quite a remarkable change took place between the contents of the first and second report for 2001. Structurally, networks 1/2001 and 2/2001 (Figures 6 and 7) look similar: they both have one central word, *Nokia*, around which most other words occurred and there were two collocational pairs outside the main structure. Two words that appeared in 2/2001, marking the change in the networks, were *decline* and *decreased*. Neither appeared in the previous time frame. In the text of the report for 1/2001 *decline* did not appear at all and decreased only appeared twice, making the sudden increase to 5 and 16 occurrences, respectively, quite noticeable. At the same time words bearing positive connotations, such as *growth* and *increased* disappeared. The connection between these events was made explicit by the fact that *sales*, a word linked to increased in the first network, was linked to *decline* in the second. After this change, the networks still retained the same structure. They always contained one major network with *Nokia* as the central node linked to about a dozen collocates. However, positive words like *increased* did not reappear.

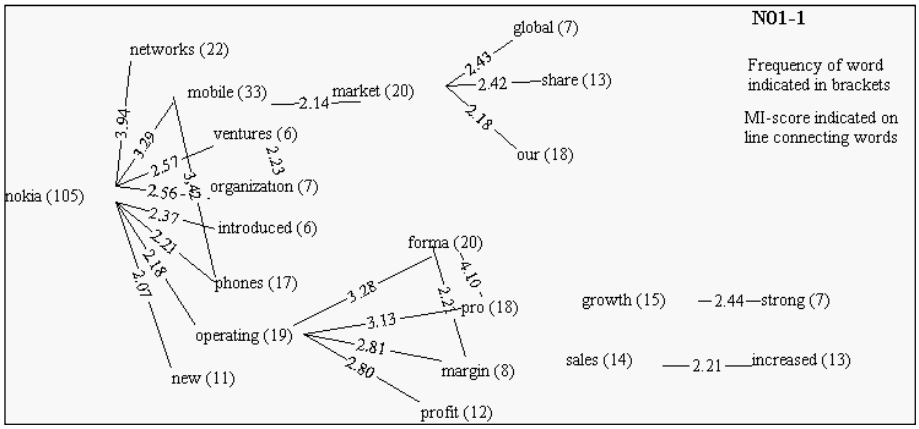

Figure 6. Collocational network of the quarterly report of Nokia in Q1/2001 (Magnusson et al. 2005)

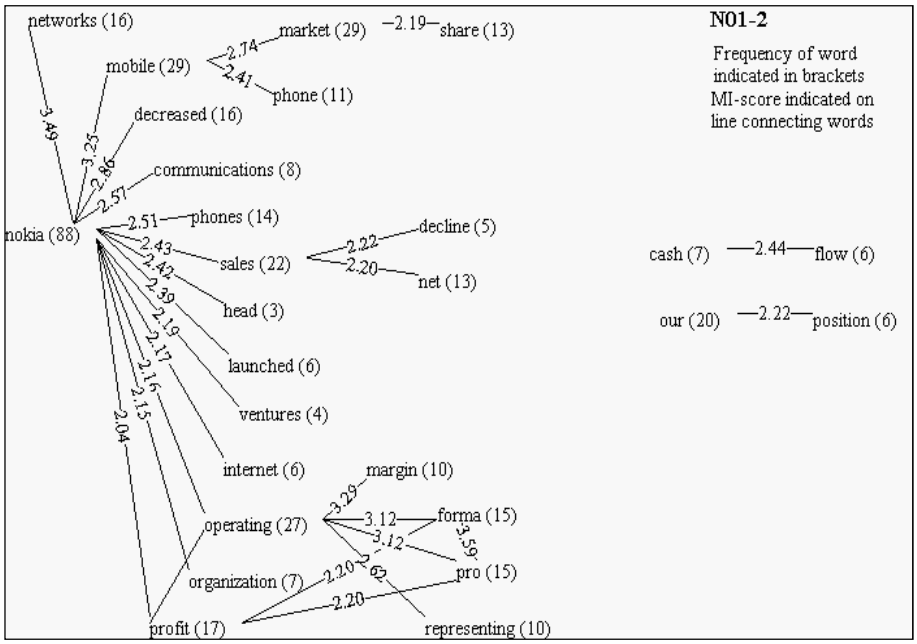

Figure 7. Collocational network of the quarterly report of Nokia in Q2/2001 (Magnusson et al. 2005)

Motorola's networks showed less uniformity than Nokia's networks. Still, for the year 2000 they resembled each other quite closely. They consisted of one main network with the word *sales* as the central node. Linked to it were words like *increased*, *higher*, *orders*, and *systems*. Interestingly, the word *lower* also appeared. A budding change could be seen in the fourth network for 2000 and particularly in the first network for 2001 (Figures 8 and 9). The main network still concentrated on *sales*, but there was also a smaller network around the word *Motorola*, which was the most frequent word in the text and was linked to the collocates *announced* and *new*. It seems that the company was trying to emphasize the announcement of new innovations. Interestingly, at the same time, positive words like *higher* and *increased* had disappeared. Network 2/2001 looks quite similar. *Motorola* was still the most frequent word, but it was now only linked to *announced*. The word *decline* had also appeared as a collocate to *sales*. In the third network for 2001, the changes continued. This looked very different from the previous ones as it did not contain any structure resembling a network,

only pairs of collocations. *Sales*, which was a central node in the previous networks, was only linked to *segment*. *Motorola* was still linked to *announced*. What made the contents of this network particularly different from previous networks, was the complete lack of words describing the financial developments, such as *increased*, *decreased*, *higher*, and *lower*.

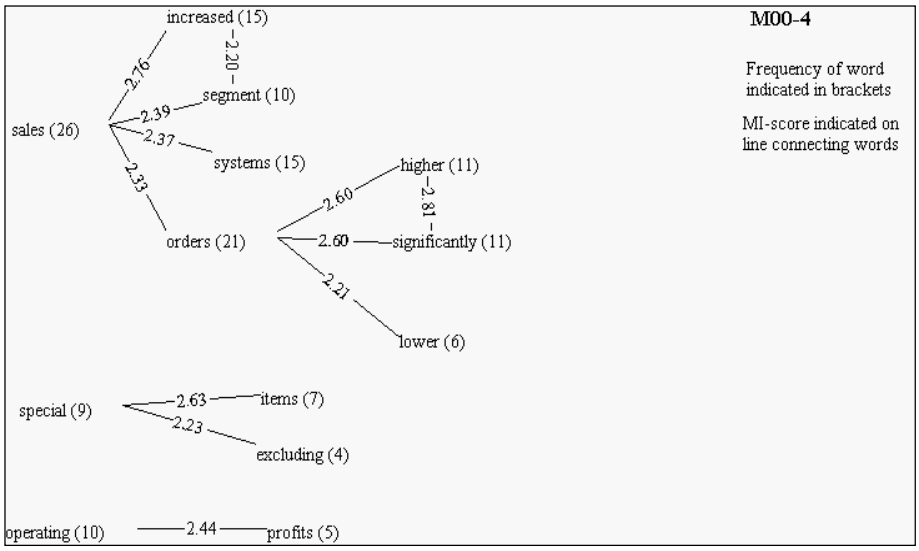

Figure 8. Collocational network of the quarterly report of Motorola in Q4/2000 (Magnusson et al. 2005)

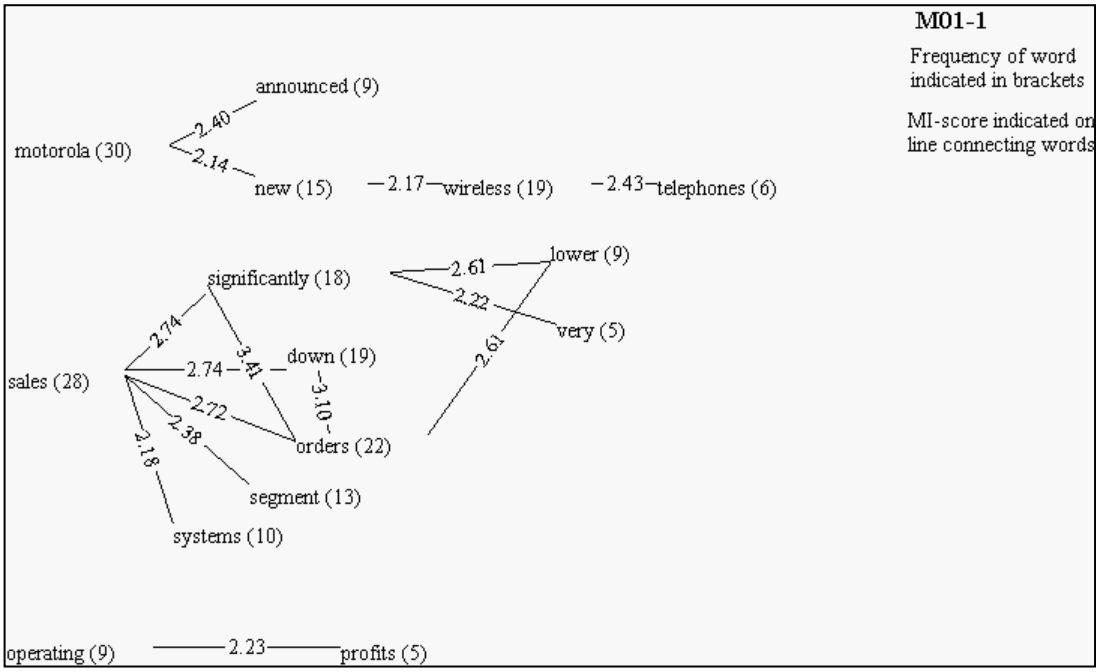

Figure 9. Collocational network of the quarterly report of Motorola in Q1/2001 (Magnusson et al. 2005)

A brief overview of the collocational networks based on Ericsson's quarterly reports showed that they never exhibited the same stability as Nokia's networks. During the period studied, both the structure and the content of the networks varied considerably. This was also obvious when looking at the texts: during this period the reports underwent several structural changes. New headings were introduced and old ones were abandoned or reorganized. A particularly remarkable change in the networks happened between the third and fourth reports for 2000 (Figures 10 and 11). Structurally, these networks were very different. There was also a significant difference between the lexical items used and the number of lexical items in the networks. Network 3/2000 started with the most frequent word, *Ericsson*, linked to five collocates. One of these,
increased, was linked to *sales*, which had four other collocates of its own. One of the collocates, *systems*, was linked to *mobile*, which had five more collocates. These linkages meant that the main network for 3/2000 consisted of three parts, connected by collocational pairs. In addition, there were several separate

collocational pairs and small networks outside the main network. The structure of network 4/2000 was very different from that of 3/2000. It consisted of a main network attached around the most frequent word, *we*, and a smaller, separate network around *operating*. *We* was a new word, and the most frequent one in network 3/2000, *Ericsson*, had disappeared: the company now referred to itself using a pronoun. In addition to these two major networks, there was one separate collocational pair, consisting of two new words, *additional* and *restructuring*; these were quite informative about Ericsson's situation. The number of words in the network was much smaller than in the previous one (33 versus 14), and the structure was much less complex. The obvious reason was that report 3/2000 consisted of approximately 3,600 words, whereas report 4/2000 had approximately 2,100.

In the next network, 1/2001, the change continued. This contained even fewer words. Now there was only one word, *expect*, connected to *we*, as opposed to five collocates in the prior network. A new addition was the collocation *efficiency program*, a term bearing obvious negative connotations to anyone acquainted with corporate jargon.

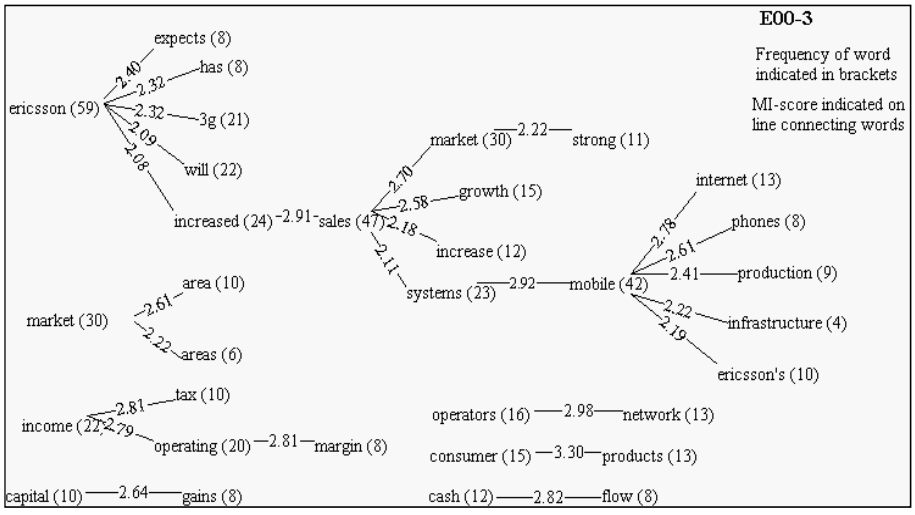

Figure 10. Collocational network of the quarterly report of Ericsson in Q3/2000 (Magnusson et al. 2005)

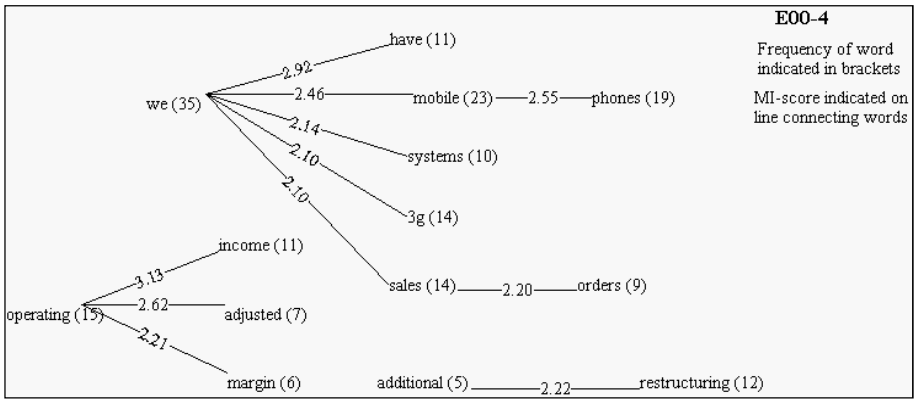

Figure 11. Collocational network of the quarterly report of Ericsson in Q4/2000 (Magnusson et al. 2005)

In summary, changes in the collocational networks seemed to take place over a period of two quarters. After a period of general stability, impending change was first heralded by a seemingly minor change, such as the appearance of a single word or word pair with a negative connotation (*expects* or *lower sales*) or the disappearance of a word or word pair indicating the continuation of previous (positive) development (*continued*). At the same time, the network contained some words with positive connotations. This can be viewed as an anticipation of change by the management of the companies. In the next quarter, the heralded change really occurred and the structure and words contained in the network changed compared to preceding quarters. If the change represented a permanent change in the environment of the company, the change became stable, as was the case for Ericsson, with the most informative collocation in quarters 3/2000–3/2001 being *efficiency program*. Table 3 contains an overview of the most significant changes in the collocational networks based on the quarterly reports.

Quarter/company	Nokia	Motorola	Ericsson
1/2000	-	-	-
2/2000	-	-	-
3/2000	-	-	-
4/2000	-	ANTICIPATION -→ lower sales	CHANGE sales growth, sales increase → additional restructuring
1/2001	ANTICIPATION continued →-	CHANGE Sales increased → significantly lower sales, sales down	- (→ efficiency program)
2/2001	CHANGE New, strong growth, increased sales → decreased, sales decline	-	-
3/2001	-	-	-
4/2001	-	-	-

Table 4. Summary of changes in the collocational networks (Magnusson et al. 2005)

After the collocational networks for the companies had been created, a visualization of their key financial figures was produced (Figure 12). On this self-organizing map, different clusters represent differences in financial performance as follows:

- Groups A1 and A2 hold the best performing companies. Profitability is very high, and solvency is good. The companies in group A2 are not quite as profitable as those in group A1, but have higher solvency and liquidity.
- Group B is slightly poorer but still very good, with good profitability and reasonable solvency.
- Groups C1 and C2 are average. Profitability and liquidity is better in group C1 than in C2, but solvency is better in group C2. Generally speaking, groups C1 and C2 are average.

- Group D is the poorest, with very low values in nearly all ratios, especially in profitability and solvency.

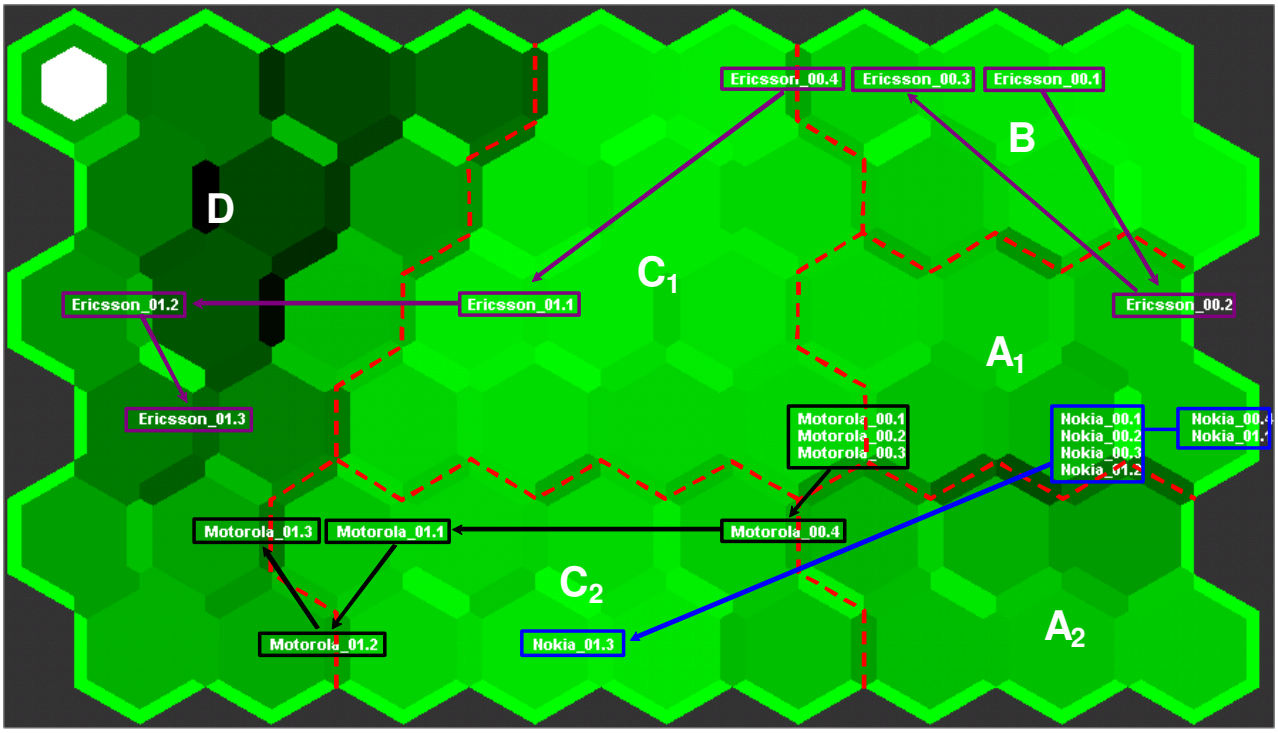

Figure 12. Movements of three telecommunications companies on the self-organizing map during 2000-2001 (Magnusson et al. 2005)

On the SOM, some major movements among the companies were observed. The financial position of Nokia was relatively stable in quarters 1/2000–2/2001 with the company firmly in the A1 group. However, a major change for Nokia occurred between quarters 2/2001 and 3/2001, when the company moved to group C2. In the third quarter's financial figures, Nokia's profitability had dropped considerably, and Nokia thus dropped from the best group into group C2. This was primarily due to defaulted loans to the Turkish telecom operator Telsim, as well as insolvency of the UK operator Dolphin.

Motorola's position was relatively stable throughout 2000, with the company in group C1 for most of the period. In Motorola's case, a change started to take place

between quarters 4/2000 and 1/2001, when the company moved to the poorer end of group C2. In 2/2001, it moved to group D.

Finally, for Ericsson the financial situation appeared relatively stable through 2000, when the company was situated in group B (with a short move to group A1). Ericsson's situation deteriorated significantly after the fourth quarter of 2000, when the company shifted first to group C1 and finally to group D. Ericsson's financial performance decreased considerably between 4/2000 and 1/2001, dropping into the lower end of group C1, indicating rapidly declining performance. During 2/2001 to 3/2001, its profitability and solvency continued to decrease, and it was obvious that the company was experiencing difficulties. As can be noted, the changes for all the companies during the studied period were to the worse, with no exception.

5.3.Results of comparing self-organizing maps and collocational networks

Results of the changes in the collocational networks (as presented in Table 4) and the significant shifts in the SOM model (as presented in Figure 12) are combined in Table 5. The table clearly shows that a significant change in the collocational network of a company's quarterly report was followed by change in the position of the company in the SOM model in the next quarter. The changes in the collocational networks were, in turn, preceded by a smaller, anticipatory change in the collocational network of the prior quarterly report compared to preceding ones.

Quarter/Company	Nokia	Motorola	Ericsson
1/2000	-	-	-
2/2000	-	-	-
3/2000	-	-	-
4/2000	-	NETWORK ANTICIPATION	NETWORK CHANGE
1/2001	NETWORK ANTICIPATION	NETWORK CHANGE	SOM SHIFT (B → C1)
2/2001	NETWORK CHANGE	SOM SHIFT (C2 → D)	SOM SHIFT (C1 → D)
3/2001	SOM SHIFT (A1 → C2)	-	-
4/2001	-	-	-

Table 5. Comparison of changes in the collocational networks and the SOM (Magnusson et al. 2005)

Looking at Nokia, a change took place in the textual material of the first and second quarters of 2001. The network for 2/2001 exhibited words such as *decline* and *decrease*. In the SOM reflecting the figures, a shift can be seen in the third quarter, when Nokia moved from an excellent performance group (A1) to the average (C2). In the case of Motorola, there was an anticipation of change in the network for 4/2000, which was significantly smaller than the previous one. The real change, however, took place in the 1/2001 network. Words like *down* appeared, and *higher* and *increased* disappeared. On the SOM, Motorola moved from the average group (C1 to C2) between the quarters 4/2000 and 1/2001, and then on to poor performance group (D) in 2/2001. For Ericsson the changes started even earlier. The networks created from Ericsson's reports for 3/2000 and 4/2000 looked completely different. The large network of the third quarter had been transformed into a much smaller one in the fourth. One of the new words appearing in 4/2000 was the clearly negative *restructuring*, whereas positive words, such as *increase* and *growth*, had disappeared. On the SOM, Ericsson moved from the good performance group (B) to the average performance group (C1) and down to the low performance group (D) during quarters 4/2000 to 2/2001. It is clear that these companies all had experienced financial difficulties

during the years 2000 and 2001, as they all shifted to worse performing groups during that time. Moreover, the companies had exhibited changes in their quarterly report texts as reflected in their collocational networks. The changes in the texts preceded the changes in the figures by approximately one quarter.

This evaluation study shows not only that collocational networks reflect changes in the underlying texts - which in the study also seem to reflect changes in the company's financial situation - but also that the combined use of quantitative and qualitative information can strengthen the management's assessment of the external environment for strategic decision making purposes. When appropriate visualization methods are used for both types of information, connections between them, such as the one in the study above, can be discovered more easily.

6 QUALITATIVE EVALUATION OF COLLOCATIONAL TOPIC NETWORKS

The evaluation discussed in Chapter 5 showed that collocational networks of financial reports reflect upcoming changes in financial figures. Still, a qualitative evaluation was needed to show whether collocational networks of financial reports also reflect changes that competitive intelligence practitioners consider to be important in these texts.

For the purposes of a qualitative evaluation, six visualizations representing the years 2003 to 2008, were created out of the annual reports of seven large telecommunications companies (see Table 6). A text file was created for each year included in the visualization by merging the seven annual reports published by the seven companies that year. A similar visualization could be made out of the reports of a single company, but it was decided that for the purposes of this evaluation, a general view of the industry would be produced. A general industry view based on the reports of all the companies would be something that all interviewees would have an opinion on and would thus be easier to discuss than developments at a single company.

Company	Country of origin	Type of document	Years included
AT&T	USA	Annual reports	2003-2008
BT	UK	Annual reports	2004-2009
France Télécom	France	Annual reports	2003-2008
Telecom Italia	Italy	Annual reports	2003-2008
Telenor	Norway	Annual reports	2003-2008
TeliaSonera	Sweden	Annual reports	2003-2008
Verizon	USA	Annual reports	2003-2008

Table 6. Texts included in the visualizations for the interviews

Next, the word *service* was selected to be the central node in the networks. Because the evaluation was made at a case company that provides value-added services to telecommunication companies, it was presumed that the interviewees would be interested in seeing the context in which the telecommunications companies discussed the concept of service. They would presumably also have opinions on how the context changed over time. This process produced the six topic networks depicted in Figure 13.

Compared to the networks used in the first evaluation (Chapter 5), some adjustments were made. The networks shown to the interviewees were considerably simpler in form, containing only the six main collocates of the central node and no further nodes connected to these. Numbers indicating word frequency and MI score were left out. This was carried out to minimize the amount of data needed to produce an interpretation of the networks.

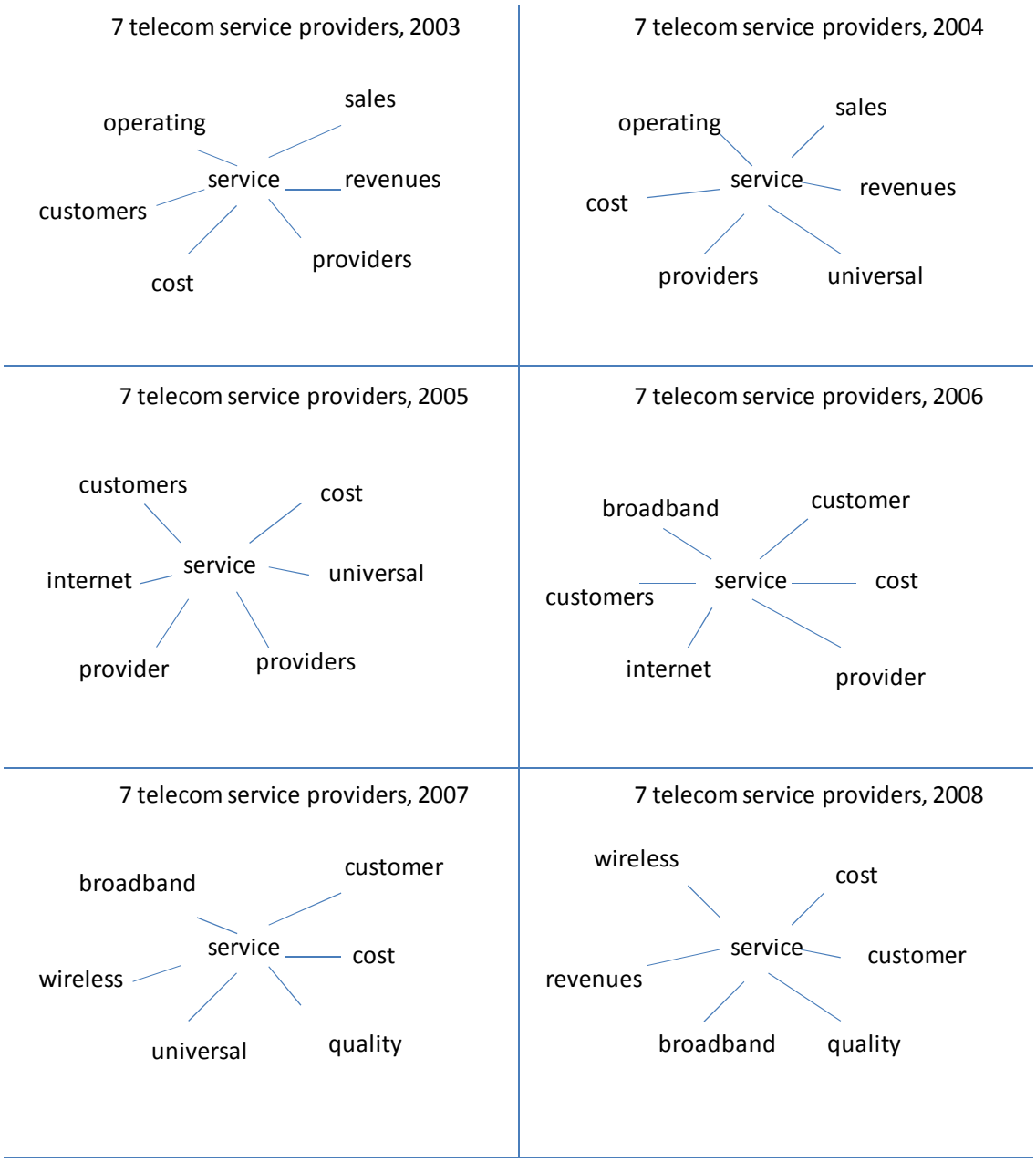

Figure 13. Topic networks for the word service, seven telecommunication companies' annual reports 2003-2008

The networks were shown to the interviewees. Their interpretations and opinions of their utility will be discussed next.

6.1.Evaluation interviews

Although still quite rare within the design science research paradigm, qualitative interviews have been used as an evaluation tool by Adomavicius et al. (2008), who used in-depth interviews with IT industry experts to evaluate a model for analyzing the IT landscape. Their paper provides an example of the appropriateness of qualitative interviews for studying complex artifacts. There are also some examples of using focus group interviews for artifact evaluation, most notably Tremblay et al (2008) as well as Gibson and Arnott (2007). The strength of qualitative interviews in evaluation lies in the fact that they allow the interviewer to explain the workings of the artifact to each interviewee and to gain more insight into possible usage scenarios as seen by the interviewees.

A challenge of using a qualitative evaluation method in design science is in measuring the outcome of the evaluation. In this thesis, the utility of the artifact is measured through answers to the following (qualitative) question: To what extent do the interviewees find that the text visualization method would help them with competitive intelligence tasks? This question encompasses a number of sub-questions: Do the interviewees think that the method effectively reflects the underlying texts? In what kind of tasks would the interviewees use the method? How would the interviewees like the method to be improved?

The evaluation in this thesis was carried out as a case study at one company. This approach was chosen because it allowed the interviewer to prepare a single set of visualizations that would be relevant to all interviewees, thus making the interviews more commensurable. In case study research, the question of the generalizability of the results is a common issue (Yin, 1989). This thesis attempts to overcome this issue by contributing with what Walsham (1995) calls “rich insight”. Here, such insights concern the limitations of text visualization in competitive intelligence, general usage scenarios, and suggestions for further

research. Research transparency (Flick, 2008) is created through detailed reporting of the background of the case company and interview subjects as well as of the interview process.

The evaluation interviews were carried out at a company that produces value-added services (VAS) for telecommunication service providers. The services produced are mainly intended for consumer broadband subscribers. Developments in the telecommunications field are thus of great importance to the case company from both a long term strategic and a day-to-day operational point-of-view. The company has some 200 telecommunications service provider partners globally, and manually analyzing the textual information that is publicly available about these partners would be an arduous task. The telecommunication companies listed in Table 6 are either existing or potential partners of the case company. They are the largest publicly listed telecommunications companies in Europe and North America that publish their financial reports in English. All visualizations in this evaluation were made out of the companies' English-language annual reports.

The eight interviewees represented the management team, the product management team and the sales department of the company. All of the interviewees had been working within the technology industry for over 10 years. Table 7 summarizes the functions and working experiences of the interviewees.

Interviewee background	
Years in information technology industry	All interviewees had 10 years or more of experience within the industry, half of them had 15 years or more
Organizational function	Management team, product management, sales department

Table 7. The backgrounds of the interviewees

The interviews were semi-structured, allowing for a conversation between interviewer and interviewee and allowing for the introduction of new topics by the interviewee. The interviewees were all asked questions on three themes:

1 Background information: the interviewees' main responsibilities within the company and the length of their work history in the company and in the industry
2 The type of external information needs in their job, the kind of information they require and the way they obtain this information, the type of information they work with (qualitative/quantitative)
3 Their interpretations of the text visualizations, i.e. the topic networks shown in Figure 13.

6.2.Interview results

Each interview lasted approximately one hour. They were recorded and notes were taken during them. In this section, the results of the interviews will be reported and discussed. First, there is an account of the interviewees' interpretations of the topic networks. Following that, there is a more detailed discussion on some themes that emerged during the interviews, concerning the use of competitive intelligence in decision making. In order not to compromise the privacy of the interviewees, individual interviews are not reported on in detail. Instead, the reporting focuses on three themes that arise from the interviews in total. They shed light on how the visualization method could be further developed to accommodate user needs. They also give some indications on why the use of text visualization methods is still very rare in a business context.

6.2.1.Interpretations of the visualizations

The interviewees were shown the list of companies included (reproduced as Table 6) and the collocational topic networks in Figure 13. At the same time, the basic workings of the visualization were explained to them in the following way: “these networks are based on annual reports by all seven telecommunications service providers, year by year. The six words around the word *service* occur in these reports in a statistically significant way together with *service*”. The interviewees where then asked to look at the networks and report on any thoughts that emerged. The interviewees were also able to ask for clarifications of the method.

After taking some minutes to study the networks, most interviewees mentioned at least two of the following observations:

- *cost* is something that seems to be mentioned in connection with service in every network. This was seen by the interviewees to reflect the fact that telecommunications companies have been struggling with cost reduction throughout this time period.
- *revenues* is a word that occurs in the first two networks, then disappears and reappears in 2008. Several interviewees noticed this, but did not quite know how to interpret it. During this period, revenues from new services have been rising, while traditional service revenues have declined.
- The earliest networks do not mention any specific service types, but in 2005 *internet* occurs for the first time. In 2006 the more specific *broadband* appears, and in 2007 *wireless* appears. The interviewees found this interesting, and believed it to reflect that these companies have become more elaborate on the different types of service they provide and now discuss the services in more detail.
- *universal* occurs in 2004, 2005 and 2007. The interviewees believed this to reflect the fact that because most of the included companies are incumbents in

their home country, they may be required by law to provide a universal service, i.e. a service that reaches all residents of that country.

- *quality* is a word that turns up in 2007, and it continues to be in the network in 2008. The interviewees saw this as an indication that fierce price competition has driven telecommunications companies to emphasize service quality in their external communication.

Generally, the interviewees considered the networks to reflect actual developments in the telecommunications service industry during the years 2003-2008. It should be noted that this method allows users to go to the original source texts, if needed, to look for explanations for the occurrence of unusual, ambiguous or otherwise particularly interesting words in the networks. In that sense, this is not a "black box" method, and user interpretations of changes in the networks can usually find support (or be disconfirmed) through the original texts.

6.2.2.Interview themes on text visualization

During the interviews, three recurring themes on the subject of text visualization began to emerge from the conversations. Themes are defined here loosely as "summary statements and explanations of what is going on" (Rubin and Rubin, 2005). These themes, which will be discussed next, are: Visualization of qualitative versus quantitative information; Uses for text visualization; Visualization for analysis versus visualization for presentation.

Theme 1: Visualization of qualitative versus quantitative information

All interviewees said that they work with both qualitative and quantitative information. They considered qualitative information to include both written and spoken verbal, non-numeric information. For many interviewees qualitative information was particularly important in forward-looking processes, it was seen as something that was essential in the context of creating new services and new

partnerships. There was, however, a strong feeling among many interviewees that quantitative information is a more important base for decision making; it was seen as concrete and unambiguous, as "facts". Several interviewees said that there were aspects of their work for which they would like to have more concrete figures, so they would not have to put so much guess work into decision making.

> *"In the end, it's all about money, which is a quantitative entity"* – Member of management team

When asked about visualizations of quantitative data, all interviewees said that they were familiar with, and frequently worked with, models such as pie charts or bar charts. The interviews show that, in contrast to this, the use of text visualization and text mining methods in a corporate context is still quite rare. None of the interviewees had encountered such methods during their working career. Thus the initial mention of text visualization did not prompt reactions in the interviewees.

To introduce the interviewees to the subject, a simple text visualization called a word cloud was shown. A word cloud is an image consisting of words in different font-sizes. Word clouds can be seen and produced e.g. on http://www.wordle.net/. They are fairly common on various websites and are sometimes used to visualize the most commonly used words on a site (the bigger the font-size, the more frequent the word is) or the most commonly used search words on the site. Several interviewees had seen these online, but none of them had worked with them. This indicates that although many potential users of text visualization methods are unfamiliar with the term *text visualization*, they may still have encountered usage of such methods online or elsewhere, which might make it easier to introduce them in a corporate setting. Still, the novelty of these methods means that they should be introduced with care. Methods for visualizing qualitative data can never be "exact" in the same sense that methods for visualizing quantitative data. When introducing these methods to a company, this should be done gradually through cases. It is important that the users feel that the

method is contributing with new insight, not just reiterating something that is already known. This was emphasized by several interviewees:

> *"(they are useful)...if they clearly show trends which aren't in a way self evident"* – Member of sales team

To improve the visual appearance of the networks, some of the interviewees suggested that different colours or different size fonts be used in the networks, to emulate the visual properties of the word cloud while still keeping the temporal insights that a series of topic networks provides.

Theme 2: Uses for text visualization

The interviewees were asked to spontaneously present situations where they would consider using such visualizations. They were allowed to mention any kind of source material for these visualization cases, and were not limited to financial reports of potential partner companies, which was the example used for evaluation purposes.

All interviewees quickly responded with various scenarios where they could see the networks to be useful. Such scenarios include an analysis of competitors' marketing materials, including differences between competitors and changes within the industry over time. An aspect that interests the interviewees is whether the actors within the industry compete by arguing for the technical superiority of their products or take a softer approach, emphasizing the advantages of these products for the end-user. A similar analysis could be conducted on product reviews in technology magazines. Several interviewees remarked that this was something that would interest the communications department as well.

> *"PR, who push out our press releases, they should have a tool like this, they should know* [what topics are being emphasized in the industry]" – Member of product management team

Financial reports, such as annual reports, were considered to be more reliable sources for information, but also to contain "old news" in the sense that they are usually backward-looking and give less indication of a company's future strategy.

For a more neutral and up-to-date view of the developments within the industry the interviewees suggested visualizing reports by third-party analysts or online technology magazines and blogs.

The interviewees' suggestions for potential uses were not restricted to external data. Some of them mentioned the potential of visualizing company internal data, such as unstructured emails from consumers to the technical support email. One interviewee also said that it would be interesting to see visualizations of internal strategy material, in order to compare them to how the case company's strategy is presented externally.

Theme 3: Visualization for analysis versus visualization for presentation

The interviewees' responses to the networks shown and their suggestions for usage scenarios show that they see the visualizations as accurate representations of the underlying texts. Furthermore, they see them as tools that could be used to make users aware of issues that would otherwise go undetected. Although this is encouraging for the further development of text visualization, there is also a risk involved: a user doing the analysis alone could put too much weight on topics that are not relevant from a decision making point of view. At the same time, more relevant topics may go unnoticed. In order to avoid such situations it would be useful for users to be able to discuss their interpretations of the visualizations in groups and compare them to information that they have obtained through other sources, such as meetings with representatives of customers, suppliers or third-party industry analysts.

Most interviewees, however, said that there are few opportunities for such meetings. Usually, they conduct their analysis alone and share the conclusions of their findings as arguments in slideshow presentations. It seems that while the interviewees could easily imagine the method for analysis, not all of them considered the visualizations unambiguous enough to be shown as part of an internal presentation.

> *"This is really a tool for analysis, you would still need to clarify the point you are trying to make with this raw data. If you showed the images as they are you would have to explain a lot"* - Member of product management team

Nevertheless, one interviewee believed that the networks, because of their novelty, could be used to capture the attention of potential customers at sales meetings, using changes in industry analyst reports over a period of time as visualization material:

> *"it could be very interesting for them* [the customers]*, they would see that they have read the same Gartner reports but haven't noticed these things"* - Member of product management team

In summary, the interviews show that the interviewees quickly discovered changes in the networks that they considered to be relevant in their work. They found the networks to reflect some actual changes that have taken place within the telecommunications industry during the years 2003-2008. In their comments they also supported the idea of using text visualization in competitive intelligence as a tool for strategic decision making.

7 CONCLUSIONS

This thesis concludes with a recapitulation of its main contribution: collocational networks as a visualization method for competitive intelligence. After that, some suggestions for further improvements to the method and to text visualization research in general will be made.

7.1. Collocational networks as a text visualization method

In this thesis, collocational networks are proposed as a simple text visualization tool for competitive intelligence practitioners working with strategic management. The networks are particularly suitable for visualizing changes in sequences of text, and alleviate the inclusion of qualitative information in an analysis of the company-external environment, where the use of quantitative information usually dominates. In contrast to many other text mining and text visualization tools, the interpretation of the networks requires no knowledge of linguistics.

Drawing on design science methodology, the networks were evaluated in this thesis using two methods, a quantitative and a qualitative one. The quantitative evaluation, carried out as a comparison between collocational networks of quarterly report texts and self-organizing maps of the financial figures of the same reports, showed that changes in the collocational network of a company's quarterly report was followed by a change in the position of the company in the self-organizing map in the next quarter.

The qualitative evaluation, carried out as interviews with competitive intelligence practitioners, showed that the interviewees considered the collocational topic networks they were shown to reflect actual developments in the telecommunications service industry during the years 2003-2008.

The results of these evaluations mean that the method can be considered to have proven its utility in text visualization. The suitability of the method for competitive intelligence purposes is further indicated by the fact that all interviewees found the visualizations to provide interesting insights and quickly thought of several possible uses for the method within the domain of competitive intelligence.

Collocational topic networks would be particularly useful as a tool for a person conducting an analysis on texts produced in the business environment of a company. Networks could in some cases also be shown to an audience, internal or external, but the need for making this type of presentation has to be carefully considered; the unfamiliarity of the visualization method to most viewers means that the visualization itself might get more attention than the argument that the presenter is trying to make using it.

The development and evaluation of the collocational topic networks in this thesis has been carried out in line with the thinking of Kroes (2002), which emphasizes the man-made artifact as an intentional object, grounded in a social context. From an evaluation point of view, this means that the qualitative interviews have focused not only on the artifact to be evaluated, but on the current competitive intelligence working practices and information visualization needs of the interviewees. This approach has also been fruitful in bringing light on why it is rare to find text mining and text visualization tools in use in companies today. Although adequate technology exists, many of these tools have not been developed in context but rather as isolated artifacts, not taking into account the backgrounds of the users and their different needs. Evaluating the collocational

networks in context also provided material for further improvements, which will be discussed next.

7.2.Suggestions for further research

A number of ideas for further improvements to the method arose during the research process, many of these during the conversations with the interviewees of the second evaluation. Some of them suggested that the networks could be made more visually appealing. The network format was well received by the interviewees, but font sizes and different colours could, for example, be used to make words that appear for the first time in the networks stand out. Different font sizes and colours could also be used for words in different predefined topic categories, such as technology, market, customers and so on.

After technical improvements, the networks should be tested in actual decision making situations, corresponding to Venable's (2006) definition of a naturalistic evaluation setting. One such situation could include a type of analysis that was suggested by one of the interviewees. In this analysis, the occurrence of certain themes, such as consumer services, in texts published by telecommunications companies, would be compared to the occurrence of the same themes in texts published by value-added service (VAS) providers, such as the case company. The aim of this comparison would be to establish which group of companies usually introduces such new themes to the market or, in the words of the interviewee, "who is driving the industry". If the comparison were to indicate that there is a significant lag between when the telecommunications companies introduce a topic and when the VAS providers introduce them, as the interviewee speculated, then the case company should use this information to its strategic advantage. It should then try to approach its partners/potential partners with topics that are relevant to them, i.e. "speak the same language", earlier than its competitors.

Another lesson learned during the research process is that in any text visualization project, be it for scholarly or for business purposes, it is important to know the intelligence needs of those involved and conduct a thorough analysis of the kind of data sources that would provide the type of information that is needed. The importance of matching the right sources with the right methods is something that is quite rarely discussed within text mining and text visualization research, and which would require more attention. Text visualization research could be further improved through the inclusion of linguistic genre research (as exemplified by Bhatia 1993) in the *knowledge base* (Hevner et al. 2004) of text visualization design research.

Within the topic of strategic management, more research is clearly needed on the actual use of qualitative and quantitative company-external information. Using methods such as discourse analysis and experimental settings, decision making situations should be studied to establish whether the addition of qualitative information is seen by practitioners to improve competitive analysis, and what kind of information is seen as particularly useful.

REFERENCES

Included are references from both the research summary and the research papers

Adomavicius, G., Bockstedt, J. C., Gupta, A., & Kauffman, R. J. (2008). Making sense of technology trends in the information technology landscape: a design science approach. *MIS Quarterly*, *32*(4), 779-809.

van Aken, J. E. (2004). Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules. *The Journal of Management Studies*, *41*(2), 219-246.

Back, B., Toivonen, J., Vanharanta, H., & Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems*, *2*(4), 249-269.

Back, B., Oosterom, G., Sere, K., & van Wezel, M. (1995). Intelligent information systems within business: bankruptcy predictions using neural networks. In *Proceedings of the 3rd European Conference on Information Systems, ECIS'95*. Athens.

Back, B., Sere, K., & Vanharanta, H. (1998). Managing complexity in large data bases using self-organizing maps. *Accounting, Management and Information Technologies*, *8*(4), 191-210.

Baskerville, R. (2008). What design science is not. *European Journal of Information Systems*, *17*(5), 441-443.

Berry-Rogghe, G. (1973). The computation of collocations and their relevance in lexical studies. In *The Computer and Literary Studies* (pp. 103-112). Edinburgh.

Bhatia, D. V. K. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems, 108*(4), 510-528.

Bose, I., & Mahapatra, R. K. (2001). Business data mining - a machine learning perspective. *Information & Management*, *39*(3), 211-225.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Clark Williams, C. (2008). Toward a taxonomy of corporate reporting strategies. *Journal of Business Communication, 45*(3), 232-264.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, *41*(6), 801-849.

Courseault Trumbach, C., Payne, D., & Kongthon, A. (2006). Technology mining for small firms: Knowledge prospecting for competitive advantage. *Technological Forecasting and Social Change*, *73*(8), 937-949.

Dutka, A. (1999). *Competitive Intelligence For The Competitive Edge*. Lincolnwood: NTC Business Books.

Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2003). Using the self-organizing map as a visualization tool in financial benchmarking. *Information Visualization*, *2*, 171-181.

Eklund, T. (2004). *The Self-Organizing Map in Financial Benchmarking*. TUCS Dissertations (Vol. 56). Turku: Turku Centre for Computer Science.

Eppler, M. (2006). A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization, 5,* 202-210.

Eppler, M., & Platts, K. (2009). Visual strategizing: the systematic use of visualization in the strategic-planning process. *Long Range Planning, 42*, 42-74.Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Commun. ACM*, *49*(9), 76-82.

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, *25*(4), 335-342.

Firth, J. (1957). *Papers in Linguistics 1934-51*. Oxford University Press, London.

Fleisher, C. S. (2003). Should the field be called "Competitive Intelligence" or something else? In *Controversies in Competitive Intelligence: The Enduring Issues* (pp. 56-69). Westport: Praeger.

Fleisher, C. S., & Bensoussan, B. (2003). Why is analysis performed so poorly and what can be done to improve it? In *Controversies in Competitive Intelligence: The Enduring Issues* (pp. 110-122). Westport: Praeger.

Fleisher, C. S., & Bensoussan, B. E. (2007). *Business and Competitive Analysis: Effective Application of New and Classic Methods*. FT Press.

Fleisher, C. S., & Blenkhorn, D. L. (2003). *Controversies in competitive intelligence*. Greenwood Publishing Group.

Flick, U. (2008). *Managing Quality in Qualitative Research*. SAGE Publications.

Gibson, M., & Arnott, D. (2007). The use of focus groups in design science research. In *ACIS 2007 Proceedings* (p. Paper 14). Presented at the Australasian Conferences on Information Systems 2007.

Goldkuhl, G. (2008). What kind of pragmatism in information systems research? Presented at the AIS SIG Prag Inaugural meeting, Paris.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman Pub Group.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). College Park, Maryland: Association for Computational Linguistics.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 87-92.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Indulska, M., & Recker, J. (2008). Design science in IS research: A literature analysis. In *Information Systems Foundations: Answering the unanswered questions about design*

research. Presented at the The 4th Biennial ANU Workshop on Information Systems Foundations, Canberra.

Järvinen, P. (2007). Action research is similar to design science. *Quality & Quantity*, *41*, 37-54.

Jönsson, S., & Lukka, K. (2006). There and back again: doing interventionist research in management accounting. In *Handbook of Management Accounting Research* (Vol. 1, pp. 373-397). Elsevier.

Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2001). *Analysing financial performance with quarterly data using self-organising maps* (No. 430). TUCS Technical report. Turku.

Kasanen, E., Lukka, K., & Siitonen, A. (1993). The constructive approach in management accounting. *Journal of Management Accounting Research*, *5*, 243-264.

Kaski, S., & Kohonen, T. (1995). Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. In *Proceedings of the Third International Conference on Neural Networks in the Capital Markets* (pp. 498-507). London.

Kendall, J. E. (1993). Good and evil in the Chairmen's 'boiler plate': an analysis of corporate visions of the 1970s. *Organization Studies*, *14*(4), 571-592.

Kim, W. C., & Mauborgne, R. (2005). *Blue Ocean Strategy: How to Create Uncontested Market Space and Make Competition Irrelevant*. Harvard Business Press.

Kiviluoto, K. (1998). Predicting bakruptcies with self-organising maps. *Neurocomputing*, *21*, 191-201.

Kjellmer, G. (1987). Aspects of English collocations. In *Corpus linguistics and beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora* (pp. 133-140).

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, *23*(1), 67-93.

Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., & Visa, A. (2002). Combining data and text mining techniques for analysing financial reports. In *Proceedings of the Eighth Americas Conference on Information Systems* (pp. 20-28). Dallas.

Kohonen, T. (1997). *Self-Organizing Maps* (second edition.). Berlin: Springer.

Kohut, G. F., & Segars, A. H. (1992). The President's letter to stockholders: an examination of corporate communication strategy. *Journal of Business Communication*, *29*(1), 7-21.

Krier M., & Zacca F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information*, *24*, 187-196.

Kroes, P. (2002). Design methodology and the nature of technical artefacts. *Design Studies*, *23*(3), 287-302.

Lehtinen, J. (1996). *Financial Ratios in an International Comparison* (No. 49). Acta Wasaensia. Vaasa.

Leong, E., Ewing, M., & Pitt, L. (2004). Analysing competitors online persuasive themes with text mining. *Marketing Intelligence & Planning*, *22*, 187-200.

Lurie, N., & Mason, C. (2007). Visual representation: implications for decision making. *Journal of Marketing, 71*, 160-177.

Länsiluoto, A., Back, B., Vanharanta, H., & Visa, A. (2001). Country specific financial trend analysis with self-organizing maps. In *Proceedings of the Tenth Annual Research Workshop on Artificial Intelligence and Emerging Technologies (AI/ET) in Accounting, Auditing and Tax* (pp. 15-23). Atlanta.

Länsiluoto, A., Back, B., Vanharanta, H., & Visa, A. (2002). Multivariable business cycle analysis with self-organizing maps – are the cycles similar? In *Proceedings of the European Conference on Accounting Information Systems*.

Magnusson, C. (2010) Improving competitive analysis with temporal text visualization. *Marketing Intelligence & Planning, 28*, 571-581.

Magnusson, C., & Vanharanta, H. (2003). Visualizing sequences of texts using collocational networks. In *Proceedings of the 3rd international conference on Machine learning and data mining in pattern recognition* (pp. 276-283). Leipzig, Germany: Springer-Verlag.

Magnusson C., Arppe A., Eklund T., Back B., Vanharanta H., & Visa A. (2005). The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management, 42*, 561-574.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251-266.

Martín-del-Brío, B., & Serrano-Cinca, C. (1993). Self-organizing neural networks for the analysis and representation of data: some financial cases. *Neural Computing and Applications*, *1*, 193-206.

Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198-207). Chicago, Illinois, USA: ACM.

Meek, G K., Roberts, C. B., & Gray, S. J. (1995) Factors influencing voluntary annual report disclosures by U.S., U.K. and continental European multinational corporations. *Journal of International Business Studies, 26*(3), 555-572.

Michalisin, M. D., & White, G. P. (2000) Validity of annual report assertions about quality: an empirical study. *The Mid-Atlantic Journal of Business, 36*, 103-122.

Mingers, J. (2001). Combining IS research methods: towards a pluralist methodology. *Information Systems Research*, *12*(3), 240-259.

Mintzberg, H. (1994). *Rise and Fall of Strategic Planning*. Free Press.

Mintzberg, H., Lampel, J., & Ahlstrand, B. (1998). *Strategy Safari: A Guided Tour Through The Wilds of Strategic Management*. Free Press.

Niehaves, B. (2007). On epistemological diversity in design science - new vistas for a design-oriented IS research? Presented at the Twenty Eighth International Conference on Information Systems, Montreal.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, Oxford.

Nunamaker, J. F., Chen, M., & Purdin, T. D. M. (1991). Systems development in information systems research. *Journal of Management Information Systems*, *7*(3), 89-106.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Orlikowski, W. J., & Baroudi, J. (1991). Studying information technology in organizations: research approaches and assumptions. *Information Systems Research*, *2*, 1-28.

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: Desperately seeking "IT" in IT research - A call to theorizing the IT artifact. *Information Systems Research*, *12*(2), 121-134.

Osborne, J. D., Stubbart, C. I., & Ramaprasad, A. (2001). Strategic groups and competitive enactment: a study of dynamic relationships between mental models and performance. *Strategic Management Journal*, *22*(5), 435-454.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech.

Phillips, M. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Elsevier Science Ltd.

Pihlanto, P. (2009). *Decision-Maker in Focus: Holistic Individual in the Theater of Consciousness.* LAP Lambert Academic Publishing..

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors.* The Free Press.

Purao, S. (2002). *Design Research in the Technology of Information Systems: Truth or Dare.* GSU Department of CIS Working Paper. Atlanta.

Purao, S., Baldwin, C., Hevner, A., Storey, V., Pries-Heje, J., Smith, B., & Zhu, Y. (2008). The sciences of design: observations on an emerging field. *Communications of the Association for Information Systems*, *23*(1), 523-546.

Rubin, H. J., & Rubin, I. S. (2005). *Qualitative interviewing: the art of hearing data*. Thousand Oaks Calif.: Sage Publications.

Saussure, F. D. (1995). *Cours de linguistique générale*. Payot.

SCIP. (2009). Glossary of terms used in competitive intelligence and knowledge management. Retrieved December 19, 2009, from http://scip.cms-plus.com/files/Resources/Prior%20Intelligence%20Glossary%2009Oct.pdf

Searle, J. R. (1995). *Construction of Social Reality*. Free Press.

Serrano-Cinca, C. (1996). Self organizing neural networks for financial diagnosis. *Decision Support Systems*, *17*(3), 227-238.

Simon, H. A. (1996). *The Sciences of the Artificial - 3rd Edition* (3rd ed.). The MIT Press.

Sinclair, J. (1991). *Corpus Concordance and Collocation*. Oxford: Oxford University Press.

Spiegler, I. (2003). Technology and knowledge: Bridging a "generating" gap. *Information and Management*, *40*(6), 533-539.

Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language, 2*, 23-55.

Sullivan, D. (2001). *Document warehousing and text mining*. New York: Wiley.

Tan, R. G. H., van den Berg, J., & van den Berg, W. (2002). Credit rating classification using self-organising maps. In *Neural Networks in Business: Techniques and Applications* (pp. 140-153). Hershey: Idea Group Publishing.

Thomas, J. (1997). Discourse in the marketplace: The making of meaning in annual reports. *Journal of Business Communication*, *34*(1), 47-66.

Tremblay, M., Hevner, A. R., & Berndt, D. J. (2008). The use of focus groups in design science research. In *Proceedings of the 3rd International Conference on Design Science Research in Information Systems and Technology* (pp. 17-37). Atlanta: Georgia State University.

Vaishnavi, V., & Kuechler, W. (2004). Design Research in Information Systems. Retrieved October 12, 2009, from http://www.isworld.org/Researchdesign/drisISworld.htm

Walle, A. H. (2000). *Qualitative Research in Intelligence and Marketing: The New Strategic Convergence*. Praeger.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research*, *3*(1), 36-59.

Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of Information Systems*, *4*(2), 74-81.

Walsham, G. (2006). Doing interpretive research. *European Journal of Information Systems*, *15*(3), 320-330.

Venable, J. (2006). A framework for design science research activities. In *Proceedings of the 2006 Information Resource Management Association Conference* (pp. 184-187). Washington, DC, USA.

Williams, G. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151-171.

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470-475.

Yang, Y., Akers, L., Klose, T., & Barcelon Yang, C. (2008).Text mining and visualization tools – impressions of emerging capabilities. *World Patent Information, 30*, 280–293

Yin, R. K. (1989). *Case Study Research: Design and Methods*. Sage Publications, Inc.

Zahra, S., & Chaples, S. (1993). Blind spots in competitive analysis. *The Academy of Management Executive*, *7*(2), 7-28.

APPENDIX: RESEARCH PAPERS

1. Magnusson Camilla & Hannu Vanharanta 2003. Visualizing sequences of texts using collocational networks. Petra Perner and Azriel Rosenfeld (Eds.) *Machine Learning and Data Mining in Pattern Recognition*. Lecture Notes in Artificial Intelligence 2734. Berlin: Springer. pp. 276-283.
   Author's contribution: Planned and conducted the research. Also wrote the research paper.

2. Magnusson Camilla, Antti Arppe, Tomas Eklund, Barbro Back, Hannu Vanharanta & Ari Visa 2005. The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management*, 42, 561-574.
   Author's contribution: Coordinated the research paper and conducted the part of the research relating to collocational networks. Also co-wrote the introduction, discussion and conclusion.

3. Magnusson Camilla 2010. Improving competitive analysis with temporal text visualization. *Marketing Intelligence & Planning, 28*, 571-581.

4. Magnusson Camilla, unpublished manuscript. A qualitative user evaluation of collocational topic networks as a text visualization method. Submitted to European Journal of Information Systems.